3.Lu R Zhao X Li J et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; (published online Jan 30.)	Genome sequences of 2019-nCoV sampled from nine patients who were among the early cases of this severe infection are almost genetically identical, which suggests very recent emergence of this virus in humans and that the outbreak was detected relatively rapidly. 2019-nCoV is most closely related to other betacoronaviruses of bat origin, indicating that these animals are the likely reservoir hosts for this emerging viral pathogen.	Figure 1A shows 8 sequences and the concensus sequence. These 8 sequences show 3 with 0 mutations, 2 with 1 mutation, 3 with 2 mutations, and none with more than 2 mutations. Based on current estimates of 1 mutation per human passage, these are at most two human-to-human transfers apart. Importantly, there is no background diversity as would be seen in two or more resevoir-to-human events. Fig 2 states strain Bat-SL-CoVZC45 is 87.6% sequence identity to the human virus, which means a difference of about 3700 mutations or over 70 years from lowest common ancestor.
4.Zhu N Zhang D Wang W et al. A novel coronavirus from patients with pneumonia in China, 2019. NEJM. 2020; (published online Jan 24.)	"more than 85% identity with a bat SARS-like CoV (bat-SL-CoVZC45, MG772933.1) genome published previously. Since the sequence identity in conserved replicase domains (ORF 1ab) is less than 90% between 2019-nCoV and other members of betacoronavirus, the 2019-nCoV — the likely causative agent of the viral pneumonia in Wuhan — is a novel betacoronavirus belonging to the	A >85% identity with a bat coronavirus means the **human and bat virus have over 70 years to LCA.**

	sarbecovirus subgenus of Coronaviridae family."	
5.Ren L Wang Y-M Wu Z-Q et al. Identification of a novel coronavirus causing severe pneumonia in humans: a descriptive study. Chin Med J. 2020; (published online Feb 11.)	All five patients have sequence homology of 99.8% to 99.9%. These isolates showed 79.0% nucleotide identity with the sequence of SARS-CoV (GenBank NC_004718) and 51.8% identity with the sequence of MERS-CoV (GenBank NC_019843). The virus is closest to a bat SARS-like CoV (SL-ZC45, GenBank MG772933) with 87.7% identity, but is in a separate clade. **Surprisingly, RNA-dependent RNA polymerase (RdRp), which is the most highly conserved sequence among different CoVs, only showed 86.3% to 86.5% nt identities with bat SL-CoV ZC45.**	Similar to reference 3 comments. Lack of conserved sequencing of the most highly conserved sequence with bat coronavirus would **suggest a non-bat source.**
6.Paraskevis D Kostaki EG Magiorkinis G Panayiotakopoulos G Tsiodras S Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol. 2020; (published online Jan 29.)	A BLAST search of 2019-nCoV middle fragment revealed no considerable similarity with any of the previously characterized corona viruses. Bat_SARS-like coronavirus sequences cluster in different positions in the tree, suggesting that they are recombinants, and thus that the 2019-nCoV and RaTG13 are not recombinants. Codon usage analyses can resolve	**The middle segment with no similarity to other corona viruses is about 40% of the entire genome. I agree SARS-CoV-2 is not a recombinant of RaTG13. I agree, codon usage analysis here supports the furin binding site insertion as having been invented de novo. A recent recombination event is not necessary for a**

	the origin of proteins with deep ancestry and insufficient phylogenetic signal or **invented de novo**. Our study rejects the hypothesis of emergence as a result of a recent recombination event. Notably, the new coronavirus provides a new lineage for almost half of its genome, with **no close genetic relationships to other viruses** within the subgenus of sarbecovirus. This genomic part comprises half of the spike region encoding a multifunctional protein responsible also for virus entry into host cells	**laboratory derived theory of origin. Statements do not advance a zoonotic origin.**
7.Benvenuto D Giovanetti M Ciccozzi A Spoto S Angeletti S Ciccozzi M The 2019-new coronavirus epidemic: evidence for virus evolution. J Med Virol. 2020; (published online Jan 29.)	The epidemic originated in Wuhan, China. A phylogenetic tree has been built using the 15 available whole genome sequences of 2019-nCoV, 12 whole genome sequences of 2019-nCoV, and 12 highly similar whole genome sequences available in gene bank (five from the severe acute respiratory syndrome, two from Middle East respiratory syndrome, and five from bat SARS-like coronavirus). >97% maximum likelihood match to Bat SARS-like virus 2015 (Fig 1) is noted. The SARS and MERS viruses are excluded as a source of SARS-CoV-2. These results do not	A 3% genome **distance from the noted bat virus to human is about 34 years** at 26 mutations per year, the in-human mutation rate. Predicted a future mutation like the D614G mutation which is more infective.

	exclude the fact that **further mutation due to positive selective pressure, led by the epidemic evolution, could favor an enhancement of pathogenicity and transmission of this novel virus.**	
8.Wan Y Shang J Graham R Baric RS Li F Receptor recognition by novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS. J Virol. 2020; (published online Jan 29.)	Based on predicted RBD-host ACE2 receptor affinities, civet, mice, and rats are fuled out as source species. Pigs, ferrets, cats, and **nonhuman primates** contain largely favorable 2019-nCoV-contacting residues in their ACE2. SARS-CoV was isolated in wild palm civets near Wuhan in 2005, and its RBD had already been well adapted to civet ACE2.	The **potential nonhuman primate ACE2 usage is noted. Consistent with a laboratory origin from VERO cells,** a monkey kidney cell line. It expresses an ACE2 that permits SARS-CoV-2 infection, making it a possible source for the virus. A common tissue culture cell line forSARS virus research.
9.US Center for Disease Control and Prevention Coronavirus disease 2019 (COVID-19) situation summary. https://www.cdc.gov/coronavirus/2019-nCoV/summary.html Date: Feb 16, 2020 Date accessed: February 8, 2020	Rarely, animal coronaviruses can infect people and then spread between people such as with MERS-CoV, SARS-CoV, and now with this new virus, named SARS-CoV-2. The SARS-CoV-2 virus is a betacoronavirus, like MERS-CoV and SARS-CoV. All three of these viruses have their origins in bats. The sequences from U.S. patients are similar to the one that China initially posted, suggesting a likely single, recent emergence of this virus from an animal reservoir.	There are no data to support these statements about bats as the source for SARS-CoV-2.

10.Andersen KG Rambaut A Lipkin WI Holmes EC Garry RF The proximal origin of SARS-CoV-2. http://virological.org/t/the-proximal-origin-of-sars-cov-2/398 Date: Feb 16, 2020 Date accessed: February 17, 2020	See Table 2.	See Table 2.
11.Bengis R Leighton F Fischer J Artois M Morner T Tate C The role of wildlife in emerging and re-emerging zoonoses. Rev Sci Tech. 2004; 23: 497-512	In one pattern, actual transmission of the pathogen to humans is a rare event but, once it has occurred, human-to-human transmission maintains the infection for some period of time or permanently. Some examples of pathogens with this pattern of transmission are human immunodeficiency virus/acquired immune deficiency syndrome, influenza A, Ebola virus and severe acute respiratory syndrome.	This 2004 paper describes the pattern of rare animal-to-human transmission followed by human-to-human spread as an example of the SARS virus. It does not address the origin of SARS-CoV-2.
12.Woolhouse ME Gowtage-Sequeria S Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005; 11: 1842-1847	Emerging and reemerging pathogens are disproportionately viruses, with 37% being RNA viruses. Emerging and reemerging pathogens more often are those with broad host ranges that often encompass several mammalian orders and even nonmammals. For pathogens that are minimally transmissible within human populations (R0 close to 0), outbreak size is determined largely by the number of introductions from the reservoir. For pathogens that are highly transmissible within human populations	This 2005 article has good general information about looking broadly for the reservoir species(s), identifies RNA viruses as a major source of human epidemics, predicts a large outbreak size for a high Ro virus, but does address the origin of SARS-CoV-2 origin.

	(R0>>1), outbreak size is determined largely by the size of the susceptible population.	
13.NASEM The National Academies of Science Engineering and Medicine of the USA. NAS, NAE, and NAM presidents' letter to the White House Office of Science and Technology Policy. https://www.nationalacademies.org/includes/NASEM%20Response%20to%20OSTP%20re%20Coronavirus_February%206,%202020.pdf Date: Feb 6, 2020 Date accessed: February 7, 2020	The closest known relative of 2019-nCoV appears to be a coronavirus identified from bat-derived samples collected in China.4 The experts informed us that additional genomic sequence data from geographically- and temporally-diverse viral samples are needed to determine the origin and evolution of the virus. Samples collected as early as possible in the outbreak in Wuhan and samples from wildlife would be particularly valuable. Understanding the driving forces behind viral evolution would help facilitate the development of more effective strategies for managing the 2019-nCoV outbreak and for preventing future outbreaks.	Agree. If additional genomic sequence data is available from geographically- and temporally-diverse viral samples are needed to determine the origin and evolution of the virus this should be made publicly available.
14.WHO Director-General's remarks at the media briefing on 2019 novel coronavirus on 8 February 2020. https://www.who.int/dg/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020 Date: Feb 8, 2020 Date accessed: February 18, 2020	A general statement about the emerging pandemic without reference to the origin of SARS-CoV-2	There is no data about the origin of the pandemic.

In November 2020 the Watchdog group, US Right-to-Know, reported the following with respect to the *Lancet* article:[59]

"Emails obtained by U.S. Right to Know show that a statement in *The Lancet* authored by 27 prominent public health scientists condemning "conspiracy theories suggesting that COVID-19 does not have a natural origin" was organized by employees of EcoHealth Alliance, a non-profit group that has received millions of dollars of U.S. taxpayer funding to genetically manipulate coronaviruses with scientists at the Wuhan Institute of Virology."

"The emails obtained via public records requests show that EcoHealth Alliance President Peter Daszak drafted the Lancet statement, and that he intended it to "not be identifiable as coming from any one organization or person" but rather to be seen as "simply a letter from leading scientists". Daszak wrote that he wanted "to avoid the appearance of a political statement."

A separate, worrisome article entitled, "Peter Daszak's EcoHealth Alliance Has Hidden Almost $40 Million In Pentagon Funding And Militarized Pandemic Science,[60]" seems to indicate a serious conflict of interest with respect to Dr. Daszak's participation in any investigations on the origin of SARS-CoV-2.

Paper 3: The March 17, 2020 article in *Nature Medicine* entitled "The proximal origin of SARS-CoV-2" by Andersen et al.[61, 62]

According to the journal, this article is in the 99th percentile (ranked 2nd) of the 312,683 tracked articles of a similar age in all journals and the 99th percentile (ranked 1st) of the 147 tracked articles of a similar age in *Nature Medicine.* The metrics also indicate it has been accessed over five million times. It is clearly the most cited paper and since its title and topic are the origin of the pandemic it clearly has an outsized influence on the topic.

The following statements form the evidence in the article of the natural origin of CoV-2:

- "While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses **predict that the interaction is not ideal** and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding. Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is **most likely the result of natural selection on a human or human-like ACE2** that permits another optimal binding solution to arise. **This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation."** [emphasis added.]

[59] https://usrtk.org/biohazards-blog/ecohealth-alliance-orchestrated-key-scientists-statement-on-natural-origin-of-sars-cov-2/
[60] https://www.independentsciencenews.org/news/peter-daszaks-ecohealth-alliance-has-hidden-almost-40-million-in-pentagon-funding/
[61] https://www.nature.com/articles/s41591-020-0820-9
[62] Two non-peer reviewed analyses are included here because they provide a nearly line-by-line analysis. They unfortunately include occasional colorful language but the content is worth noting: https://harvardtothebighouse.com/2020/03/19/china-owns-nature-magazines-ass-debunking-the-proximal-origin-of-sars-cov-2-claiming-covid-19-wasnt-from-a-lab/ ; https://www.youtube.com/watch?v=HnnSCMb8Nds4

- A later analysis of over 3800 possible substitutions of amino acids in a 200 amino acid receptor binding region, much larger than the small, selective region referred to in this paper, shows that CoV-2 is 99.5% optimized for binding to the ACE-2 receptor. This near perfect binding has never been seen before in a recent interspecies transmission jump.

- "Polybasic cleavage sites have not been observed in related 'lineage B' betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have those sites and predicted O-linked glycans. Given the level of genetic variation in the spike, **it is likely that SARS-CoV-2-like viruses with partial or full polybasic cleavage sites will be discovered in other species.**" [emphasis added.]

  - As of the writing of this manuscript no other lineage B (sarbecovirus) has been found to have a furin site. In addition, the furin site of CoV-2 has the unusual -CGG-CGG- codon dimer, which has never been seen in an analysis of 58 other sarbecoviruses, that is, 580,000 codons. Since recombination between subgenera of beta coronaviruses is rare, or unknown, there is no source for the CGG-CGG dimer via a natural recombination event.

- "The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals."

  - It is curious why the above statement did not lead to a hypothesis somewhere in the article about a similar mechanism on CoV-2, a clear indication of a laboratory origin.

- "It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-CoV-like coronavirus."

  - This conclusory statement is unsupported my evidence.

- "Furthermore, if genetic manipulation had been performed, one of the several reverse-genetic systems available for betacoronaviruses would **probably have been used**. However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone." [emphasis added.]

  - There is no explanation for why a prior backbone would necessarily be used. All synthetic biology chimera coronaviruses created in the past as published in prior papers have each used a unique backbone with no particular pattern in backbone selection. Each backbone was selected for the particular needs of those current experiments. This non-repeating prior pattern of reverse-genetic systems makes the above statement untenable. And with 16,000+ reported coronavirus specimens at the WIV it entirely reasonable a non-published virus could have been used.

- "Natural selection in an animal host before zoonotic transfer. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for binding to human ACE2, **an animal host would probably have to have a high population density (to allow natural selection to proceed efficiently**) and an ACE2-encoding gene that is similar to the human ortholog." [emphasis added.]

    - The paragraph discusses the pangolin as the possible intermediate host but at the time of this manuscript the coronavirus data from pangolins has been discredited. This author agrees with statement that selection of the two unique features of CoV-2 require a high population density of the animal host. Of course, in the laboratory the animal hosts for either *in vitro* cell culture experiments or in animal experiments are a single species at high density.

- Natural selection in humans following zoonotic transfer. "It is possible that a progenitor of SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during **undetected human-to-human transmission**. Once acquired, these adaptations would enable the pandemic to take off and produce a sufficiently large cluster of cases to trigger the surveillance system that detected it." [emphasis added.]

- "Studies of banked human samples could provide information on whether such cryptic spread has occurred. Further serological studies should be conducted to determine the extent of prior human exposure to SARS-CoV-2."

    - As will be shown in later sections, this prior undetected human-to-human transmission would be evident in archived specimens from before the fall of 2019. In both SARS-CoV-1 and MERS, this prior seroconversion averaged about 0.6% with almost 5% among workers exposed to the intermediate hosts. At the time of the writing of this manuscript, in limited sampling of archived specimens there has been no seroconversion detected. The author believes there are thousands of archived specimens from Wuhan taken in the fall of 2019 and these should be immediately examined for evidence of seroconversion. Since finding seroconversion among these specimens would be strong evidence for a zoonotic origin and not a laboratory accident, the absence of any information from China on this important evidence is hard to understand.

- Selection during passage. "Basic research involving passage of bat SARS-CoV-like coronaviruses in cell culture and/or animal models has been ongoing for many years in biosafety level 2 laboratories across the world, and there are documented instances of laboratory escapes of SARS-CoV. We must therefore examine the possibility of an inadvertent laboratory release of SARS-CoV-2."

- "In theory, it is possible that SARS-CoV-2 acquired RBD mutations during adaptation to passage in cell culture, as has been observed in studies of SARS-CoV."

- "New polybasic cleavage sites have been observed only after prolonged passage of low-pathogenicity avian influenza virus in vitro or in vivo. Furthermore, a hypothetical generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with very high genetic similarity, **which has not been described.** Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to those of humans, **but such work has also not previously been described**." [emphasis added.]

    - The authors correctly describe a method for CoV-2 to have been generated in the laboratory and then dismiss it because the work has not been published previously. As active scientists themselves, the authors must know how disingenuous this sounds. Almost by definition elite scientists, like Dr. Shi of the WIV, work in secret until the publication of any given line of research. As the say, the absence of evidence cannot be used as evidence of its absence.

    - A peer-reviewed paper[63] entitled, "Might SARS-CoV-2 Have Arisen via Serial Passage through an Animal Host or Cell Culture? A potential explanation for much of the novel coronavirus' distinctive genome," provides a compelling argument that serial passage in the laboratory might indeed have been the manner in which CoV-2 acquired many of its devastating traits.

- "Although the **evidence shows that SARS-CoV-2 is not a purposefully manipulated virus**, it is currently impossible to prove or disprove the other theories of its origin described here. However, **since we observed all notable SARS-CoV-2 features, including the optimized RBD and polybasic cleavage site, in related coronaviruses in nature, we do not believe that any type of laboratory-based scenario is plausible**." [emphasis added.]

    - This author could identify no prior evidence in the paper to warrant saying it is not a purposefully manipulated virus. There is also no evidence that would point to a purposely manipulated virus.

    - The evidence in the paper shows that no prior zoonotic interspecies transmission has ever had an RBD as optimized as the CoV-2 RBD for the human ACE2. The evidence also shows that there is no natural source for the polybasic cleavage site (PCS). No other member of the subgenera to which CoV-2 belongs has a PCS. Since these are the only coronaviruses from which recombination could supply a polybasic cleavage site, the data in this paper refutes the natural origin.

[63] https://onlinelibrary.wiley.com/doi/full/10.1002/bies.202000091

- The belief statement concerning a laboratory-based scenario would be closer to the evidence if it was professed with, "despite evidence which is consistent with a laboratory-based scenario."

Based on the author's analysis of the paper, the following email was sent to the lead author:

Gmail Steven Quay, MD, PhD [redacted]

SARS-CoV-2 origin
[redacted]

Steven Quay, MD, PhD [redacted] Mon, May 25, 2020 at 7:14 PM
To: [redacted]

Dr. Andersen-

I read with interest your paper titled, The proximal origin of SARS-CoV-2, in which you conclude this is a natural, zoonotic-sourced infection. Three different approaches of analysis that I have done do not support this conclusion. Can you comment please:

1. The furin cleavage site insert has the unusual codon usage for the RR dimer of CGG-CGG. As you probably know, the frequency of this codon usage in the SARS-CoV-2 genome is 0.09. So having two next to each other is not likely as a random event. As support for the unlikeliness of these codons, using GISAID data, by March there is evidence of the third G being mutated out for either A or T at three-times the rate of the background mutation rate, 26/year from Nextstrain.org. Since codon usage in coronaviruses are not greatly influenced by the host it resides in, this means a jump to humans must have been in a host which did not have even a few months history with the virus, otherwise the terminal G would have been purified out. On the other hand, most laboratory use optimized codon primers and kits use CGG routinely; including in primers in published papers from the Wuhan Institute of Virology. So a laboratory source for a gain-of-function furin cleavage site would probably use these codons.

2. In the over 16,000 genomes in GISAID there is not one example of posterior diversity. With MERS, 93% of sequenced genomes did not pass through the index case but represented separate reservoir host to human jumps and this was apparent within 60 days of the index case. They collectively showed the most recent common ancestors among themselves was over 12 months before the index case. With SARS-CoV-2 it is acting like a 'pure culture' growth from the index case outward with no evidence of a reservoir host in the background. This would be the case for a laboratory acquired infection.

3. If you use a map of Wuhan and overlay the first four hospitals that saw cases with a map of the Metro system, you see that the hospitals straddle Line 2, which runs approximately east to west, carries 1,000,000 people a day, and is the Metro line with stops closest to both the Wuhan Institute of Virology and the original wet market that was considered an early source for the infection. There are 11 Metro lines in Wuhan, hundreds of stops on those lines, and over a dozen hospitals spread out over the city. I am working with a UCLA statistician to perform tests about the probability of this being simply an accident of statistics but the gestalt is, it does not look like a chance occurrence. But it is consistent with someone getting infected in the lab, riding Line 2 for a few days, and off you go.

It might be a truism to say that the six proven cases of laboratory derived SARS escapes occurred in big cities, Beijing, Singapore, or Taipei where the labs are located. But if you follow it with the fact that MERS and SARS, both proven as true zoonotic sources, on other hand began in rural settings in China and the Middle East, respectively. I am not sure why this obvious correlation was not at least pointed out in your paper and then addressed with a cogent argument.

I look forward to hearing your thoughts.

Regards, Steve

Soon after this email was written Dr. Andersen blocked the author from following his Twitter account. A reply to the above email was never received.

Conclusion. Three high visibility papers were published between January and May 202 which purported to settle the question of the origin of SARS-CoV-2 as a zoonotic transmission and not a laboratory accident. The analysis above concludes that these papers are not persuasive. The

author has elected to not use evidence within these papers to change the prior likelihood of a zoonotic versus laboratory origin. They are presented here as neutral evidence that supports neither theory.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

Evidence. SARS-like infections among employees of the Wuhan Institute of Virology in the fall of 2019

The State Department of the United States issued the following statement on January 15, 2021[64]:

"1. Illnesses inside the Wuhan Institute of Virology (WIV):

- The U.S. government has reason to believe that several researchers inside the WIV became sick in autumn 2019, before the first identified case of the outbreak, with symptoms consistent with both COVID-19 and common seasonal illnesses. This raises questions about the credibility of WIV senior researcher Shi Zhengli's public claim that there was "zero infection" among the WIV's staff and students of SARS-CoV-2 or SARS-related viruses."

There is no additional evidence to support either parties position in the above statement. The U.S. Government statement would be considered hearsay in a court of law and probably not admissible. The veracity of Dr. Shi's statement above could be called into question due to other inconsistencies in some of her testimony, as reported elsewhere in this document.

At this time, the above evidence cannot be used to change the likelihood of either theory about the origin of SARS-CoV-2. The statement is kept within this analysis with the hope that in the future new information will come to light that could make this evidence a useful addition to the overall analysis.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

[64] https://2017-2021.state.gov/fact-sheet-activity-at-the-wuhan-institute-of-virology//index.html

Evidence. A Bayesian Analysis of one aspect of the SARS-CoV-2 origin, where the first recorded outbreak occurred, increases the probability of a laboratory origin.

Introduction. The two competing hypotheses of the origin of SARS-CoV-2 as a natural, zoonotic spillover event versus a laboratory-acquired infection (LAI) or other laboratory accident each had supporting evidence from the very beginning of the pandemic.

On the one hand, about 40% of early patients with COVID-19 had an association with the Hunan Seafood Market in Wuhan. Since this mirrored SARS-CoV-1, where markets selling civet cats were determined to be the origin of that human epidemic, the natural origin hypothesis seemed logical. The Chinese CDC have now ruled out the market as a source for the outbreak.

On the other hand, the laboratory origin hypothesis also had an early beginning with the fact that the outbreak began adjacent to the only high security, BSL-4 laboratory in all of China, and one of the top coronavirus research centers in the world, was the Wuhan Institute of Virology (WIV). The hospitals of the first COVID patients were very close to the WIV.

This evidence statement is taken from an article applying a Bayesian analysis to the hypothesis that the proximal origin of SARS-CoV-2 was an uncontrolled[65] release from a laboratory using, as evidence, one aspect of the SARS-CoV-2 origin story — where the first recorded outbreak occurred.[66]

Hypothesis: The first recorded outbreak of SARS-CoV-2 in the human population occurred in a city that is also home to a virology laboratory that actively performs research on closely related viruses.

In this case, the city is Wuhan, and the virology laboratory is run by the Wuhan Institute of Virology.

Analysis. This analysis set the likelihood of a laboratory escape (the prior probability the hypothesis was true) at three values, 0.01%, 0.1%, and 1.0%. The second term was the conditional probability of the evidence, given that the hypothesis is actually false. This was set at 0.01. Finally, the third term was the conditional probability of the evidence, given the hypothesis is true. This was set, biasing to the natural origin, at 0.71.

Results. The paper provides the three-by-three cube of results for the three parameters of interest.

The ardent sceptic's probability begins at 0.01% and the revised estimate is no more than 0.05% or 5/10000. It applies to someone who was initially very skeptical about a lab origin (0.01% probability), who believes there is no more than 51% chance that an uncontrolled release of a highly contagious disease would lead to a local outbreak, and who thinks there was at least a

[65] By using the term uncontrolled release, the author was specifically excluding from consideration the possibility that the pathogen was deliberately released from the laboratory.
[66] https://jonseymour.medium.com/a-bayesian-analysis-of-one-aspect-of-the-sars-cov-2-origin-story-where-the-first-recorded-1fbdcbea0a2b

10% chance that a natural outbreak of a virus native to Yunnan would have occurred in Wuhan before any place else.

On the other extreme, is the ardent believer who started with at least a 1% belief in a laboratory outbreak, is 100% certain that an uncontrolled laboratory release would result in a local outbreak and believes that the probability that a natural outbreak of a virus native to Yunnan would occur in Wuhan before any place else is less than 0.1%. The ardent believer's revised belief is that the probability that the Wuhan outbreak was caused by an uncontrolled laboratory release changes from 1% to at least 91%.

In the center, is the so-called "central" observer who accepts that the central values for each of the parameter ranges are reasonable estimates of the true values of the probability being estimated. The central observer started with an initially skeptical belief in the hypothesis of 0.1%, believes that average citizen in Wuhan was a likely as any other citizen of China to be the initial vector of the virus into the human population and believes that there is no more or less than a 71% chance that an uncontrolled release from a laboratory of a highly contagious pathogen such as SARS-CoV-2 would result in a local outbreak as opposed to an outbreak in some other location. The central observer's revised belief in the hypothesis is 6.8%. If the central observer began with a 1% belief in a laboratory origin, this analysis would change that to 41.8%.

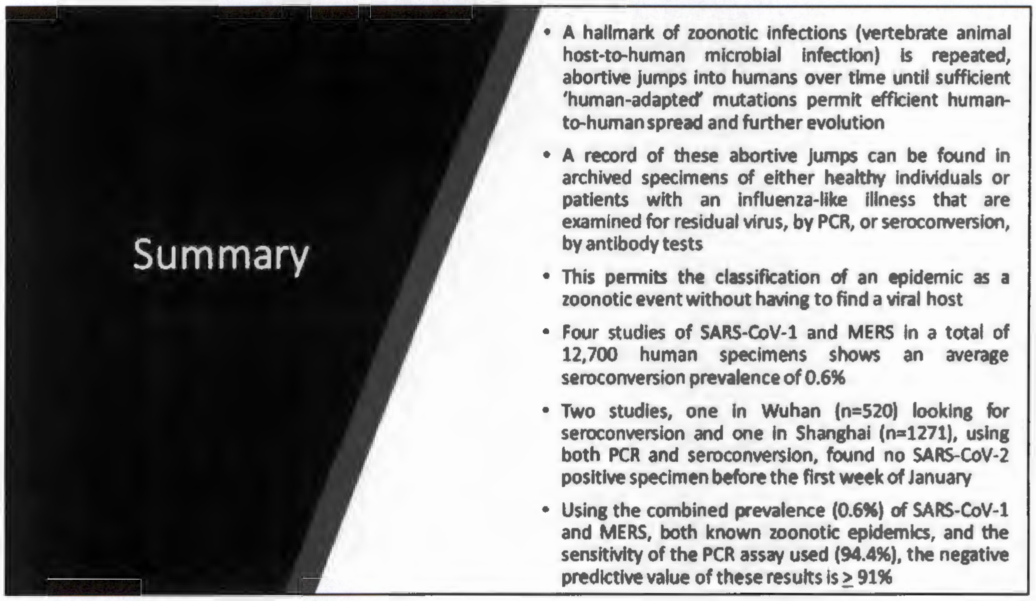
Conclusion. For purposes of this analysis and to be as conservative as possible, the assumptions will be that there is at least a 1% prior belief in a laboratory outbreak (because that was our starting probabilities), but there is no more than a 51% chance that an uncontrolled release of a highly contagious disease would lead to a local outbreak, and that there was at least a 10% chance that a natural outbreak of a virus native to Yunnan would have occurred in Wuhan before any place else. Using these assumptions, the initial likelihood of a 1% laboratory origin changes to 4.9%.

Starting likelihood from initial state: Zoonotic origin (98.8%) and laboratory origin (1.2%)

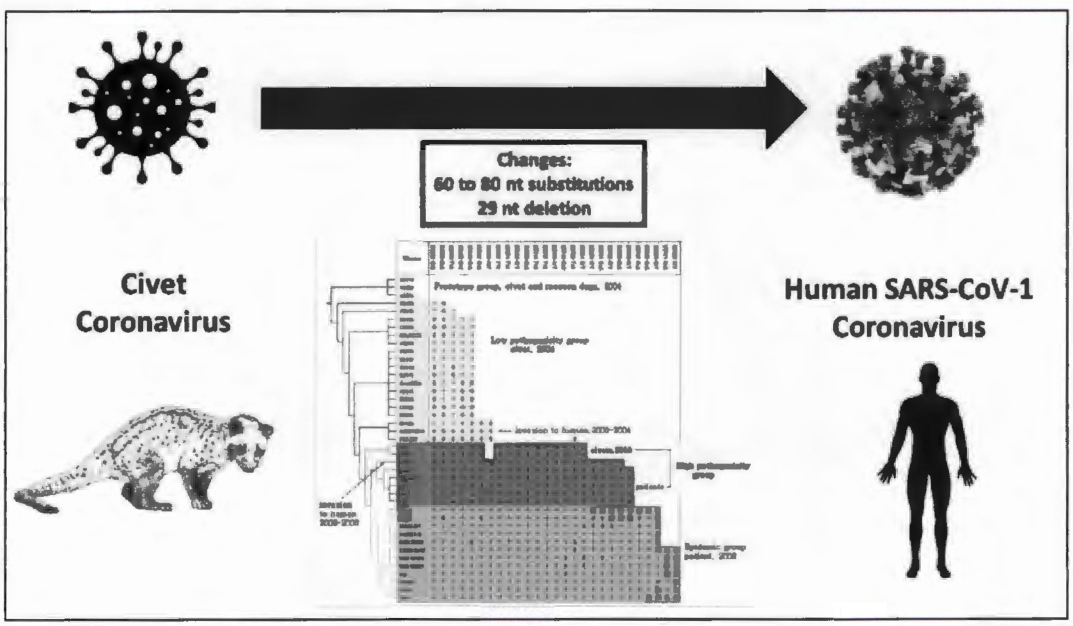
Adjusted likelihood: Zoonotic origin (95.1%) and laboratory origin (4.9%)

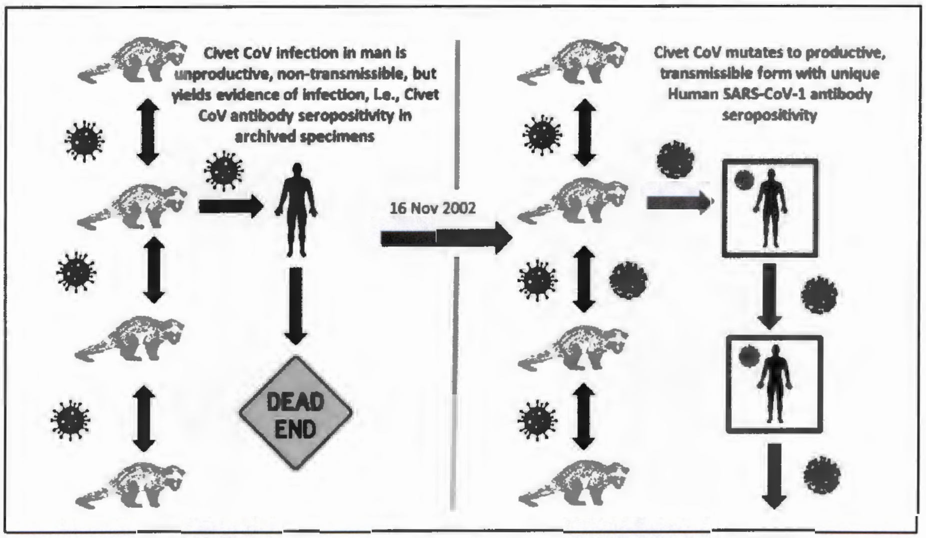
Evidence: Lack of seroconversion in Wuhan and Shanghai. Summary of evidence:

- A hallmark of zoonotic infections (vertebrate animal host-to-human microbial infection) is repeated, abortive jumps into humans over time until sufficient 'human-adapted' mutations permit efficient human-to-human spread and further evolution

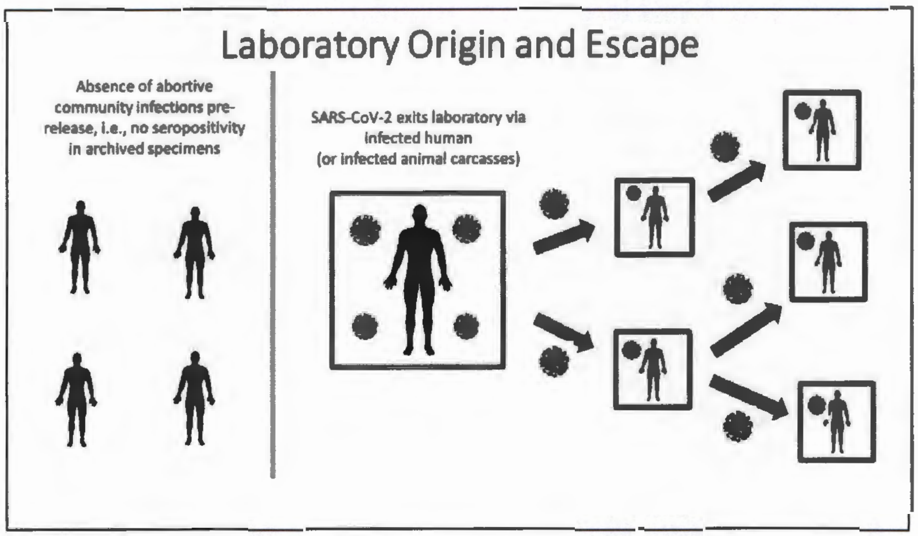


- A record of these abortive jumps can be found in archived specimens of either healthy individuals or patients with an influenza-like illness that are examined for residual virus, by PCR, or seroconversion, by antibody tests





- This permits the classification of an epidemic as a zoonotic event without having to find a viral host
- A laboratory accident is a situation in which there are no prior exposures within the human population as shown in the Figure below:



- Four studies of SARS-CoV-1 and MERS in a total of 12,700 human specimens shows an average seroconversion prevalence of 0.6%

SARS-related Virus Predating SARS Outbreak, Hong Kong

SARS-CoV-1 began in fall of 2002 in southern China

Patient Population	Serum samples collected in May 2001 from 938 healthy adults in Hong Kong	48 confirmed SARS patients diagnosed in February and March 2003 in Guangdong
Civet CoV > SARS-CoV-1 Seropositivity	13	0
SARS-CoV-1 > Civet CoV Seropositivity	4	48
Total	17 out of 938 = 1.8%	48 out of 48 = 100%

Pre-epidemic seroprevalence in the adult community

Prevalence is 0.6% for SARS-CoV-1 and MERS in 12,700 specimens

Epidemic	Nature of the Study	Seropositivity	Reference
SARS-CoV-1	Archived specimens from healthy adults in Hong Kong collected two years before CoV-1 were tested for Ab to civet or human CoV	17/938	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322899/
MERS	Archived human sera collected in 2011 was tested for MERS-CoV S1-specific antibodies by ELISA	1/90	https://www.sciencedirect.com/science/article/pii/S1876034120300010#fig0010
SARS-CoV-1	Serum specimens collected from military recruits from the People's Republic of China in 2002 were tested for SARS-CoV-1 antibodies.	11/1621	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1074388/
MERS	Between Dec 1, 2012, and Dec 1, 2013, 10,009 individual serum samples were tested for anti-MERS-CoV antibodies in regions without cases.	15/10,009	https://pubmed.ncbi.nlm.nih.gov/25863564/
SARS-CoV-1	Serum samples that were collected from 42 individuals during 2001-2002, before the SARS outbreak, and tested for IgG antibody against SARS-CoV.	28/42	https://arxiv.org/ftp/arxiv/papers/1305/1305.2659.pdf

Pre-epidemic seroprevalence in MERS shepherds and slaughterhouse workers is higher

Prevalence is 2.3% (2/87) in shepherds and 3.6% (5/140) in slaughterhouse workers

Reference: https://pubmed.ncbi.nlm.nih.gov/25863564/

- Two studies, one in Wuhan (n=520) looking for seroconversion and one in Shanghai (n=1271), using both PCR and seroconversion, found no SARS-CoV-2 positive specimen before the first week of January

Pre-epidemic seroconversion has never been seen for SARS-CoV-2

Epidemic	Nature of the Study	Seropositivity	References
SARS-CoV-2	RNA PCR from 1271 nasopharyngeal swab samples, as well as the prevalence of IgM, IgG, and total antibodies against SARS-CoV-2 in 357 matched serum samples collected from hospitalized patients with influenza-like illness between 1 December 2018 and 31 March 2020 in Shanghai Ruijin Hospital. First positive was January 25, 2020.	0/1271	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7473166/pdf/TEMI_9_1785952.pdf
SARS-CoV-2	Re-analysed 5200 throat swabs collected from patients in Wuhan with influenza-like-illness from 6 October 2019 to week one January 2020 and found no positive specimens for SARS-CoV-2 RNA by quantitative PCR.	0/520	https://www.nature.com/articles/s41564-020-0713-1
	CoV-2 Studies Combined	**0/1791**	**Probability is one in 14,881**

- Using the combined prevalence (0.6%) of SARS-CoV-1 and MERS, both known zoonotic epidemics, and the sensitivity of the PCR assay used (94.4%), the negative predictive value of these results is $\geq$ 91%

Negative Predictive Value of SARS-CoV-2 PCR Test

BioGerm PCR Test has a sensitivity of 94.4%

SARS & MERS Seroconversion	**0.60%**
PCR Sensitivity	**94.40%**
Negative Predictive Value Calculation	**<0.6/(0.6 + 0.054)**
Negative Predictive Value	**≥91%**

Here, the negative predictive value (NPV) represents the probability that a CoV-2 is not a zoonosis, given the negative seroconversion findings.

Subjective Discount Factor: 90% (a one in 10 chance this is wrong). This is a subjective value.

The change in origin likelihoods from this evidence and the calculations are shown in the Text-Table below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.951	0.049
Negative predictive value of lack of seroconversion	0.91	
Reduced by 90% Subjective Discount Factor	0.91 x 0.9 = 0.82	
Impact of this evidence	Reduces the likelihood of ZO by 82/18 or 4.6-fold. For every 100 tests, a true ZO would be seen 18 times and a non-ZO would be seen 82 times	
Impact of evidence calculation	0.951/4.6 = 0.207	
Normalize this step of analysis	0.207/(0.207 + 0.049) = 0.809	0.049/(0.207 + 0.049) = 0.191

Adjusted likelihood: Zoonotic origin (80.9%) and laboratory origin (19.1%)

Evidence: Lack of posterior diversity for SARS-CoV-2 compared to MERS and SARS-CoV-1

- The earliest stages of human CoV-1 and MERS infections were characterized by viral genome base diversity as expected for multiple, independent jumps from a large and diverse intermediate host population into humans.
- Combining MERS and CoV-1 studies, out of the earliest 255 human infections in which virus genome sequences are available, 137 could not be rooted in a prior human-to-human infection and so are attributed to an independent intermediate host-to-human infection.[67]
- That is about 54% non-human-to-human transmission.
- On the other hand, Ralph Baric has written[68] that CoV-2 is different: "SARS-CoV-2 probably emerged from bats, and early strains identified in Wuhan, China, showed limited genetic diversity, which suggests that the virus **may have been introduced from a single source**." [emphasis added.]
- With CoV-2, there are 249 viral genomes in GISAID from Hubei province, where Wuhan is located, collected between Dec 24, 2019 and Mar 29, 2020.
- From Dec 24, 2019 to November 2020, there are 1001 genomes sequenced from all of China and 198,862 worldwide.
- For CoV-2, every single genome sequence is rooted in the first sequence from the PLA Hospital in Wuhan.
- Not one case of posterior diversity.
- Using the frequency of non-rooted genome diversity seen with MERS and CoV-1, about 50:50 or a coin toss, the probability that CoV-2 is a zoonotic pandemic with 0/249 genomes is the chance of tossing a coin 249 times and getting heads every time!
- Mathematically that is nonexistent; specifically, one in 10 with 84 zeros.
- Since Wuhan had approximately 500,000 cases during the time interval of this sampling, the potential sampling error of testing only 249/500,000 or 0.05% is significant. This sampling error, while large, is unable to obliterate the overwhelming odds that this did not arise from an intermediate host in Wuhan.
- Therefore, to permit continued evidence analysis, this finding will be set at the boundary of customary statistical significance, a p-value of 0.05 or a 1 in 20 likelihood that this is zoonotic.

---

[67] https://elifesciences.org/articles/31257#abstract ; https://www.researchgate.net/publication/225726653_Molecular_phylogeny_of_coronaviruses_including_human_SARS-CoV ; https://science.sciencemag.org/content/300/5624/1394/tab-pdf ; https://pubmed.ncbi.nlm.nih.gov/14585636/ ; https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.016378-0?crawler=true ; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7118731/

[68] https://www.nejm.org/doi/10.1056/NEJMcibr2032888

Detailed explanation

A fundamental difference between a laboratory and a non-laboratory acquired zoonotic disease, the imprint of phylogenetic diversity through pre-human spread within the source population, can be examined by the posterior diversity of human cases with no *a priori* knowledge of an intermediate host.

MERS. The MERS epidemic has been documented to have arisen from the initial jump from bats to camels, a three-to-five-year expansion within the camel population in which mutational diversity arose by random mistakes, and then a jump into humans. This model of spread predicts that there would, at some point, be additional jumps from other camels into other patients, and a pattern of "posterior diversity," would be found in the human specimens. If the COVID-19 pandemic arose by a similar mechanism the same pattern would be seen. The following Text-Table contains such data.

Phylogenetic Feature	MERS	SARS-CoV-2
Posteriority Diversity	28/30 (93%)	0
No Posteriority Diversity	2/30 (7%)	7666
Time from first patient to first example of posterior diversity	About 60 days	None at >120 days
Depth of posterior diversity to first patient	>365 days	None

The study of MERS noted above was published in 2013 in Lancet[69] in an article entitled, "Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study." Thirty specimens were used in the analysis. The features of a camel-to-human zoonotic epidemic are easily identified. Specimens taken within sixty days of the first patient, "Patient Zero," began to show a background diversity that could not be traced back through Patient Zero. The analysis of all thirty, in fact, documented that 93% were transmitted directly from the camel intermediate reservoir. And looking only at the "background" diversity permitted a calculation of the last common ancestor for the spread within the camel population of over 365 days.

A study of SARS-CoV-2[70] available May 5, 2020 and entitled, "Emergence of genomic diversity and recurrent mutations in SARS-CoV-2," looked at 7666 patient specimens from around the world for phylogenetic diversity. The authors state: "There is a robust temporal signal in the data, captured by a statistically significant correlation between sampling dates and 'root-to-tip' distances for the 7666 SARS-CoV-2 ($R^2 = 0.20$, $p < .001$). Such positive association between sampling time and evolution is expected to arise in the presence of measurable evolution over the timeframe over which the genetic data was collected." This conclusion also argues against a

[69] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898949/
[70] https://www.sciencedirect.com/science/article/pii/S1567134820301829

MERS-like pattern of posterior diversity. In fact, the 95% upper bound for the probability of no posterior diversity being seen in SARS-CoV-2, given the data in MERS, is 3.9 x 10^{-4}.

The finding of posterior diversity in MERS was seen quickly, that is, within 60 days of the first patient and in only 30 specimens. In this study of COVID-19 the cutoff date of the 7666 specimens was April 19, 2020 or approximately 140 days after the first documented case. The lack of posterior diversity in COVID-19 at a much later date than what was seen with MERS also argues against a non-laboratory source for this pandemic.

A useful avenue of future research for those working to find an animal source for COVID-19 would be new mathematical models or statistical methods that might find a "hidden" signal of posterior diversity in the current data set which shows none. And given access to the unprecedented quantity of human data for COVID-19 which can be mined via bioinformatics, efforts to find the "missing link" in the wild through search and sample should be a second priority to mining the human specimen data set.

SARS-CoV-1. A similar pattern of clinical cases that do not show a common ancestor in the human population but instead is evidence of posterior diversity is shown in the Text-Table on the left for SARS-CoV-1[71] compared to CoV-2 on the right[72]. SARS-CoV-1 shows clusters of cases in humans that are connected only by phylogenetic branches that reach back in time (all of the branches inside the purple box. This is because of the extensive mutational background created while being in the intermediate host, the civet. With CoV-2 on the right, every clinical case descends from the first clinical case, in the 19A clade. There are no background mutations to account for. I will show elsewhere that the first Clade A patient was at the PLA Hospital about 3 km from the WIV.

[71] https://pubmed.ncbi.nlm.nih.gov/14585636/
[72] https://nextstrain.org/

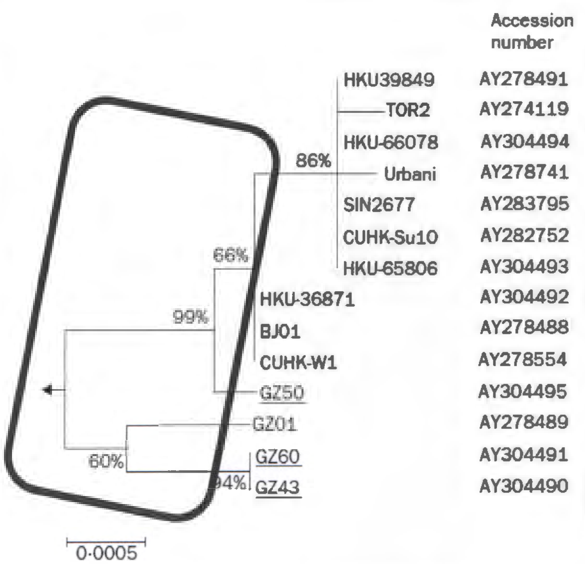
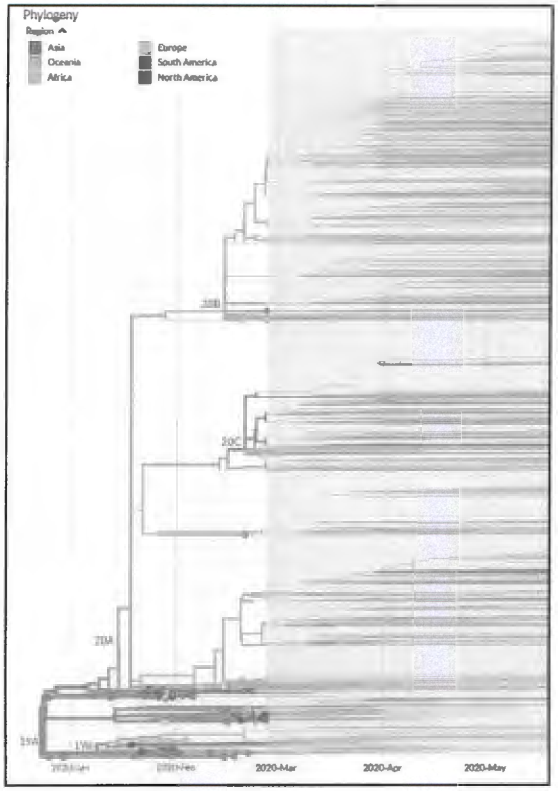
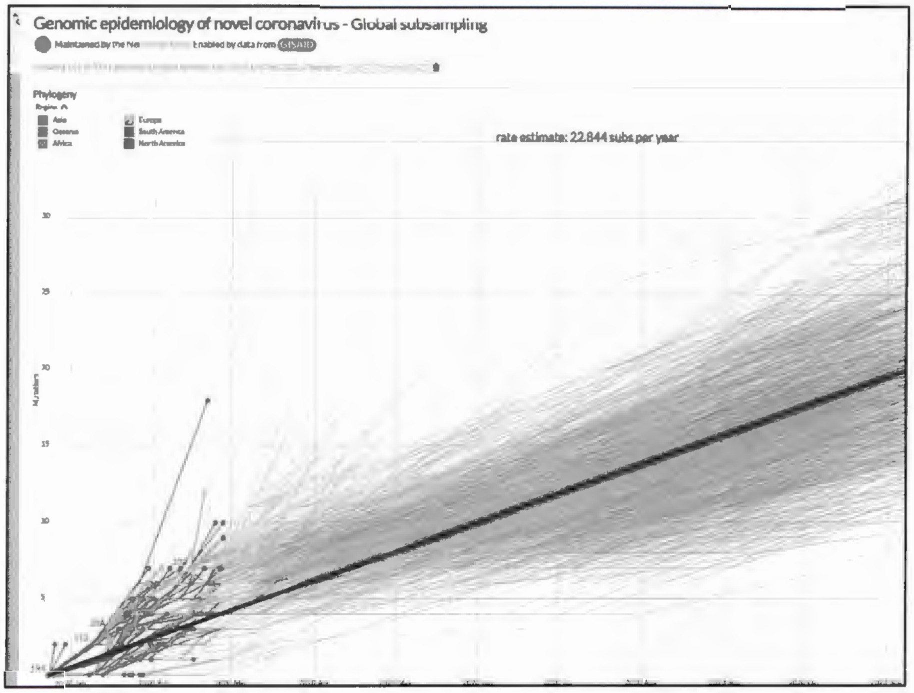


Figure 4: **Phylogenetic analysis of nucleotide acid sequence of spike gene of SARS CoV viruses**
Bootstrap values are shown as a percentage. The scale bar shows genetic distance estimated using Kimura's two parameter substitution model.[14] The nucleotide sequences of representative SARS CoV S genes (S gene coding region residue, 3765 bp) were analysed. Viruses sequenced in this study are underlined, and the other sequences used in the analysis can be accessed in GenBank with accession numbers as shown.



Given the rate of mutations of 22.8 per year for CoV-2 as shown in the Nextstrain graph below and a sequencing accuracy of about two calls per genome, CoV-2 could not have spent more than a few weeks in an intermediate host before a pattern of background mutations would be identified as posterior diversity. In the laboratory a pure culture on a single genome is used and the CoV-2 pattern is most consistent with a single pure culture infection a first human.

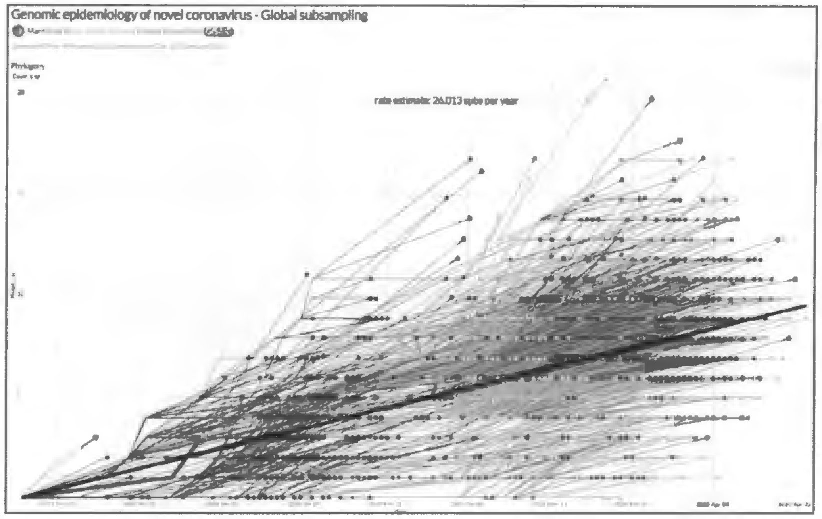


Non-zoonotic evolution. In a hypothetical in which there was a singular event in which one genetically pure virus infected one person and then the epidemic grow the development of the genetic diversity would have a clear, identifiable pattern: every new mutation would only appear on a background of the previous mutations.

The mutations in this virus are literally a personal tag. The general mutation rate leads to one mutation per patient. So, by definition, Patient Zero will have just one mutation. And then the 2-4 people that patient passes it to will have that mutation and then will add a new one, and so on. As time goes by two things happen: each patient gets a new mutation of their own and they pass on all the mutations of the past.

Since the virus has 29,900 nt and the mutation rate, as shown in this graph prepared by NextStrain is 26 mutations per year, there is very little chance a mutation will appear and then later get undone. By carefully going back in time, it is possible to literally name each person at each generation by the one (on average) new mutation they have and all of those that went before.

This graph of mutations on the Y-axis shows them gradually increasing and the color coding shows where they came from. In this infection, they only came from a previous patient and from the next previous patient and so on.



A NextStrain graphic.

How is that different from MERS, which was passed from camels to humans in a true zoonotic process?

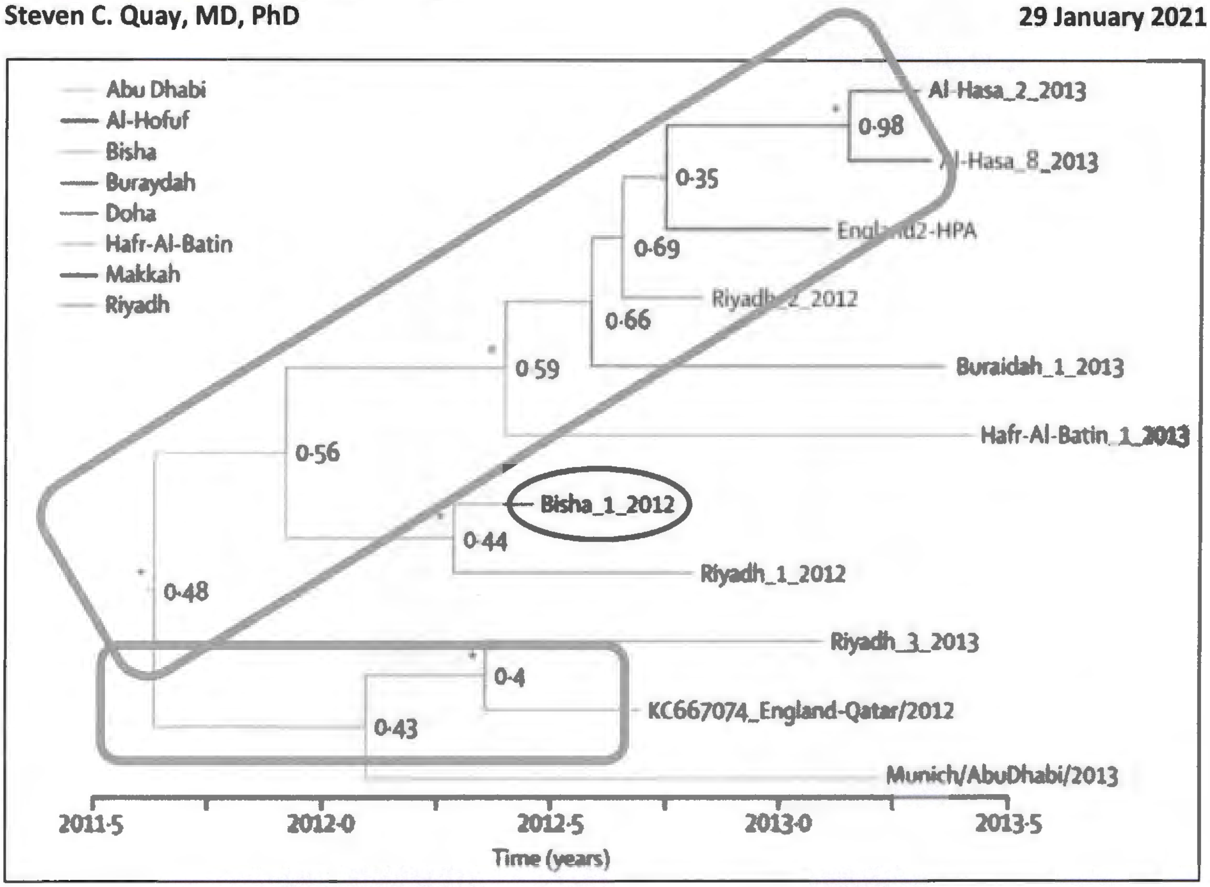
In a true zoonotic spread to humans there is usually an initiating species (in MERS it is bats), and then an intermediate species (in MERS it is camels), and then it moves to humans, either because of a new "enabling mutation" or for a non-domestic species, a chance encounter, and Source Zero and Patient Zero meet, and a cross species event occurs. But "Source Zero" doesn't stop there with one infection in one human; the virus also transmits itself vertically into the intermediate species. Source Zero also creates a vertical infection in the camels. Whether it is mild or not doesn't matter. The new human jumping gene is moving into a very diverse population of viruses, who have themselves been evolving since the first bat to camel transmission.

What is the outcome in terms of a test to show this is happening?

The diversity of the virus in humans becomes great, and the spots where the mutations occur don't match up to MERS Patient Zero like they do in COVID-19. In MERS, the virus in Patient Zero and the virus in a later infection are not direct descendants but cousins and only descended from an earlier virus that spent time in another camel population, collecting random mutations until it got the one it needed to infect humans, and then it begins again.

The chart below, from Lancet. 2013 Dec 14; 382(9909): 1993–2002, shows just how this works. The patient at Bisha is the earliest case in this chart (Patient Zero in the red circle). But notice, no other case comes from that patient. The viruses have such a diverse genetic background they appear to only be related to the Bisha virus with a posterior timeline of about one year. Their background is in the green boxes and it skips Patient Zero.

.



Even without knowing that camels are the zoonotic source for MERS, this data, from clinical sample only and without any field work in cave or camels, is all you need to know that this arose in the wild.

A paper just appeared with this analysis for a region of China and the posterior genomic diversity indicated a single starting point on December 1, 2019 for all cases. There was no posterior diversity. At this point with over 322,000 full genomes sequenced[73] and all showing an additive pattern of mutations and with none showing background diversity before the known appearance in Wuhan, the only conclusion is that there is no reservoir of genetic diversity.

On January 26, 2020 in an article in *Science* written by Jon Cohen, Kristian Andersen, an evolutionary biologist at the Scripps Research Institute who had analyzed sequences of CoV-2 to try to clarify its origin said: "The scenario of somebody being infected outside the market and then later bringing it to the market is one of the three scenarios we have considered that is still consistent with the data. It's entirely plausible given our current data and knowledge."

The negative predictive value of finding no posterior diversity in CoV-2 with 322,000 total infections sequenced, over 1000 in China, is 95%

Subjective Discount Factor: 95% (a one in 20 chance this is wrong)

[73] https://www.gisaid.org/

Below is the impact of the pack of posterior diversity on the likelihood of a zoonotic versus laboratory origin

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.809	0.191
Negative predictive value of lack of posterior diversity	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.809/9 = 0.085	
Normalize this step of analysis	0.085/(0.085 + 0.191) = 0.308	0.191/(0.085 + 0.191) = 0.692

Adjusted likelihood: Zoonotic origin (30.8%) and laboratory origin (69.2%)

Evidence: Opportunity.

The Wuhan Institute of Virology has publicly disclosed that by 2017 it had developed the techniques to collect novel coronaviruses, systematically modify the receptor binding domain to improve binding or alter zoonotic tropism and transmission, insert a furin site to permit human cell infection, make chimera and synthetic viruses, perform experiments in humanized mice, and optimize the ORF8 gene to increase human cell death (apoptosis).

Wuhan Institute of Virology scientists maps RBD and then takes a civet coronavirus that won't infect human cells, changes two amino acids in the receptor binding domain & it infects human cells.[74]

中国科技论文在线 http://www.paper.edu.cn
THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 280, No. 33, Issue of August 19, pp. 29588–29595, 2005
© 2005 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in U.S.A.

Identification of Two Critical Amino Acid Residues of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Its Variation in Zoonotic Tropism Transition via a Double Substitution Strategy*

Received for publication, January 19, 2005, and in revised form, June 16, 2005
Published, JBC Papers in Press, June 24, 2005, DOI 10.1074/jbc.M500662200

Xiu-Xia Qu,[a,b] Pei Hao,[b,c] Xi-Jun Song,[a,b] Si-Ming Jiang,[a,b] Yan-Xia Liu,[a] Pei-Gang Wang,[a] Xi Rao,[a] Huai-Dong Song,[d] Sheng-Yue Wang,[e] Yu Zuo,[a] Ai-Hua Zheng,[a] Min Luo,[a] Hua-Lin Wang,[f] Fei Deng,[f] Han-Zhong Wang,[f] Zhi-Hong Hu,[f] Ming-Xiao Ding,[a] Guo-Ping Zhao,[c,e,g,h] and Hong-Kui Deng[a,i]

From the [a]Department of Cell Biology and Genetics, College of Life Sciences, Peking University, Beijing 100871, the [c]Bioinformation Center/institute of Plant Physiology and Ecology/Health Science Center, Shanghai institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, the [d]State Key Laboratory for Medical Genomics/Pôle Sino-Francais de Recherche en Sciences du Vivant et Génomique, Ruijin Hospital Affiliated with the Shanghai Second Medical University, Shanghai 200025, the [e]Chinese National Human Genome Center, 250 Bi Bo Road, Zhang Jiang High Tech Park, Shanghai 201203, the [f]State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, 430071, and the [g]State Key Laboratory of Genetic Engineering/Department of Microbiology, School of Life Science, Fudan University, Shanghai 200433, China

Baric & Shi at WIV take bat coronavirus that won't infect human cells, change S746R to add an ARG at S1/S2 site to make furin-like cleavage site, & the new coronavirus infects human cells.[75]

Baric & Shi of WIV create completely synthetic coronavirus from bat spike & mouse adapted backbone that no treatment, monoclonal antibody, or vaccine will touch.[76]

- "Using the SARS-CoV reverse genetics system2, we generated and characterized a chimeric virus expressing the spike of bat coronavirus SHC014 in a mouse-adapted SARS-CoV backbone.

- The results indicate that group 2b viruses encoding the SHC014 spike in a wild-type backbone can efficiently use multiple orthologs of the SARS receptor human angiotensin

[74] http://www.paper.edu.cn/scholar/showpdf/NUT2kN0INTT0gxeQh
[75] https://jvi.asm.org/content/jvi/89/17/9119.full.pdf
[76] https://pubmed.ncbi.nlm.nih.gov/26552008/

converting enzyme II (ACE2), replicate efficiently in primary human airway cells and achieve in vitro titers equivalent to epidemic strains of SARS-CoV.

- Additionally, in vivo experiments demonstrate replication of the chimeric virus in **mouse lung with notable pathogenesis.**

- Evaluation of available SARS-based immune-therapeutic and prophylactic modalities revealed poor efficacy; both monoclonal antibody and vaccine approaches failed to neutralize and protect from infection with CoVs using the novel spike protein.

- On the basis of these findings, we **synthetically re-derived an infectious full-length** SHC014 recombinant virus and demonstrate robust viral replication both in vitro and in vivo."

This study was conducted, with permission, during the gain of function moratorium put in place by NIH in 2014:

"These studies were initiated before the US Government Deliberative Process Research Funding Pause on Selected Gain-of-Function Research Involving Influenza, MERS and SARS Viruses (http://www.phe.gov/s3/dualuse/Documents/gain-of-function.pdf). This paper has been reviewed by the funding agency, the NIH. Continuation of these studies was requested, and this has been approved by the NIH."

Drs. Daszak and Shi becomes world's expert on ORF8 induced apoptosis by CoVs in human cells (HeLa) & maximizing lethality.[77]

> The full-length ORF8 protein of SARS-CoV is a luminal endoplasmic reticulum (ER) membrane-associated protein that induces the activation of ATF6, an ER stress-regulated transcription factor that activates the transcription of ER chaperones involved in protein folding [35]. We amplified the ORF8 genes of Rf1, Rf4092 and WIV1, which represent three different genotypes of bat SARSr-CoV ORF8 (S3C Fig), and constructed the expression plasmids. All of the three ORF8 proteins transiently expressed in HeLa cells can stimulate the ATF6-dependent transcription. Among them, the WIV1 ORF8, which is highly divergent from the SARS-CoV ORF8, exhibited the strongest activation. The results indicate that the variants of bat SARSr-CoV ORF8 proteins may play a role in modulating ER stress by activating the ATF6 pathway. In addition, the ORF8a protein of SARS-CoV from the later phase has been demonstrated to induce apoptosis [28]. In this study, we have found that the ORF8a protein of the newly identified SARSr-CoV Rs4084, which contained an 8-aa insertion compared with the SARS-CoV ORF8a, significantly triggered apoptosis in 293T cells as well.

This paper also demonstrates the collection of 64 novel bat coronaviruses from caves in southern China, including Yunnan where Dr. Shi has said is the location of the bat ancestor of CoV-2.

This evidence is necessary for a laboratory origin hypothesis in which genetic manipulation to create CoV-2 is a precursor to a laboratory accident. However, it does not per se, provide

[77] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5708621/

increased weight in favor of a laboratory origin. It is however provided here to be a guide for the kinds of investigations to be conducted if access to the WIV records is ever provided.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (30.8%) and laboratory origin (69.2%)

Evidence and Motive for laboratory furin site insertion:

A key to infectivity of coronaviruses is the addition, in nature or the laboratory, of a furin cleavage site (FCS) at the S1/S2 junction of the Spike Protein.

Furin cleavage sites (FCS) have been widely understood to be important for many viral infections, including HIV, influenza, and others. It has also been widely understood before now that lineage B coronaviruses do not have FCS.

It was therefore surprising when an examination of SARS-CoV-2 Spike Protein found an insertion of a 12-nt, 4-AA sequence near the junction of the S1/S2 subunits which creates a furin site that is essential to human infectivity and transmission. As expected from previous work, no lineage B (sarbecovirus) coronavirus has this feature. This is the most difficult "molecular fingerprint" of SARS-CoV-2 to explain having been acquired in the wild and for that reason there are no even passingly feasible theories.

One database of whole genome sequences of 386 coronaviruses was devoid of furin cleavage sites.[78] Another database of 2956 genomes of sarbecovirus strains sequences shows that none have a furin site.[79] This is a highly significant finding with a probability that sarbecovirus has a furin site in the wild of one in about 985.[80]

It has been known since 1994 that viral glycoproteins can be cleaved by secreted proteases, including furin.[81] Even before that, in 1992, it was known the peptide sequence R-X-K/R-R in surface glycoproteins was required for avian influenza viruses of Serotype H7 pathogenesis.[82] The first paper using furin inhibitors to define a role for an FCS in coronavirus-cell fusion was published in 2004.[83]

Since that time, it has become common practice to insert FCS during laboratory gain-of-function experiments to increase infectivity. The following Text-Table illustrates the scope of just a few of the experiments conducted, with the hyperlink to the paper in column one.

URL for Paper	Title of Paper
One	Characterization of a panel of insertion mutants in human cytomegalovirus glycoprotein B.
Two	Insertion of the two cleavage sites of the respiratory syncytial virus fusion protein in Sendai virus fusion protein leads to enhanced cell-cell fusion and a decreased dependency on the HN attachment protein for activity.

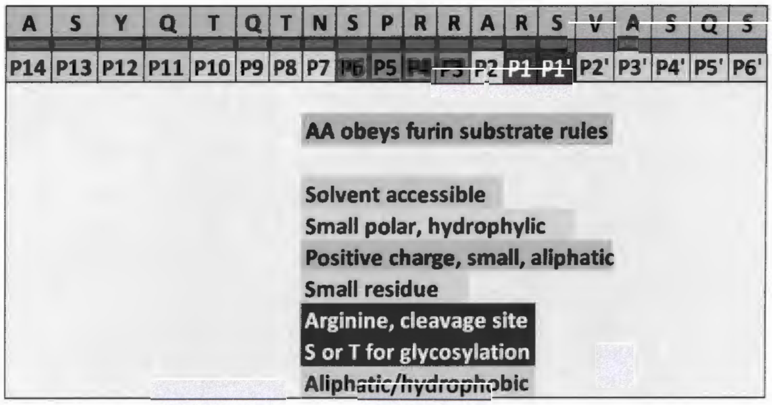
[78] https://academic.oup.com/bioinformatics/article/36/11/3552/5766118
[79] https://academic.oup.com/database/advance-article/doi/10.1093/database/baaa070/5909701
[80] When a series of samples are taken and none produce the result expected, the probability that this is a false negative finding can be estimated by taking the number of samples and dividing by three. Here, 2956 sarbecoviruses without a single furin site is a probability of one in 2956/3 or 985.
[81] https://www.ncbi.nlm.nih.gov/pubmed/8162439
[82] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7172898/pdf/main.pdf
[83] https://www.ncbi.nlm.nih.gov/pubmed/15141003

Three	Recombinant Sendai viruses expressing fusion proteins with two furin cleavage sites mimic the syncytial and receptor-independent infection properties of respiratory syncytial virus.
Four	Amino acid substitutions and an insertion in the spike glycoprotein extend the host range of the murine coronavirus MHV-A59
Five	Induction of IL-8 release in lung cells via activator protein-1 by recombinant baculovirus displaying severe acute respiratory syndrome-coronavirus spike proteins: identification of two functional regions.
Six	Coronaviruses as vectors: stability of foreign gene expression.
Seven	Experimental infection of a US spike-insertion deletion porcine epidemic diarrhea virus in conventional nursing piglets and cross-protection to the original US PEDV infection.
Eight	Minimum Determinants of Transmissible Gastroenteritis Virus Enteric Tropism Are Located in the N-Terminus of Spike Protein.
Nine	Reverse genetics with a full-length infectious cDNA of the Middle East respiratory syndrome coronavirus.
Ten	Construction of a non-infectious SARS coronavirus replicon for application in drug screening and analysis of viral protein function
Eleven	A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo.

The creation in the wild of a coronavirus FCS that is used as an example of what might have happened in SARS-CoV-2 is uninformative. In this case, a strain of influenza, in which a new polybasic site appears spontaneously leads to increased infectivity and lethality,[84] was reported by Tse *et al*. 2014. The mechanism of the FCS acquisition in this paper is an RNA polymerase dependent stuttering at a small, constrained loop in which one or more A nt were inserted, removing the strain in the loop and inserting an AAA codon which represents the basic amino acid lysine. No such method exists for the insertion of arginine, the amino acid in the CoV-2 furin site that needs to be created.

The insert generates a canonical 20 AA furin site sequence. In 2011 Tian et al.[85] published an analysis of 126 furin cleavage sites from three species: mammals, bacteria and viruses. The analysis showed that when the furin sites are recorded as a 20-residue motif, a canonical structure emerges. It includes one core cationic region (eight amino acids, P6–P2′) and two flanking solvent accessible regions (eight amino acids, P7–P14, and four amino acids, P3′–P6′).

[84] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3911587/
[85] https://www.nature.com/articles/srep00261



This figure above shows the 20-AA of the furin motif in SARS-CoV-2 (in green) with the P14 to P6' AA positions marked with the cleavage site being the amide bond between P1-R and the P1' residue. The motif is color coded with the requirements (in most cases, except for the positively charged AA requirements, most position requirements can be relaxed).

With the insertion, all 20 residues obey the rules as established by Tian. Since there are 20^4 different 4-AA peptides or 160,000 choices, it is remarkable that the 4 AA insert created a sequence that contained a small or cationic AA (8 AA/20 qualify), a cationic AA (3/20), another cationic AA (3/20), and a small AA (5/20) in that order. In fact, there are only 360 or the total or about 0.2% of all four amino acid inserts that would be expected to follow the exact rules for furin substrates. Of course, given the increase in infectivity SARS-CoV-2 has over other coronaviruses that do not have a well-designed furin cleavage site, selection pressure would drive this rare mutational event once it happened randomly. It would also be a likely choice for a laboratory designed furin cleavage site created *de novo*.

Based on the evidence that there are no furin cleavage sites in 2956 sarbecovirus (beta coronavirus) genome sequences[86], the likelihood that CoV-2 acquired the furin site from a wild sarbecovirus is one in 985 or 0.001. Because this is highly significant, we will use the conservative rule established in the beginning and use a likelihood of 0.05 for this evidence.

Subjective Discount Factor. 95% confidence (only a one in 20 chance this is wrong). Below is the calculation of the Bayesian adjustment.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.308	0.692
Negative predictive value of a lack of furin sites in sarbecovirus genomes	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.308/9 = 0.034	
Normalize this step of analysis	0.034/(0.034 + 0.692) = 0.047	0.692/(0.692 + 0.034) = 0.953

Adjusted likelihood. Zoonotic origin (4.7%), laboratory origin (95.3%).

86 https://academic.oup.com/database/advance-article/doi/10.1093/database/baaa070/5909701

Evidence: Codon usage can distinguish insertion events in the wild from those created in the laboratory.

Not only is the insertion of an FCS peptide unique among lineage B coronaviruses, the nt sequence used for the process is more broadly unique among coronaviruses in general, regardless of lineage:

-CCT-**CGG-CGG**-GCA-

I will now use synonymous codon bias methods to try to inform the question of the origin of SARS-CoV-2.

Because of the redundancy of the genetic code, more than one 3-nt sequence specifies any given amino acid. For example, there are six codons that specify arginine, R. The frequencies with which such synonymous codons are used are unequal and have coevolved with the cell's translation machinery to avoid excessive use of suboptimal codons that often correspond to rare or otherwise disadvantaged tRNAs. This results in a phenomenon termed "synonymous codon bias," which varies greatly between evolutionarily distant species and possibly even between different tissues in the same species.

Decades of research has identified that all life forms, viruses, bacteria, and humans alike, use the codons in a signature pattern of frequency which can be used to identify a particular sequence of RNA or DNA as human or non-human; viral or non-viral.

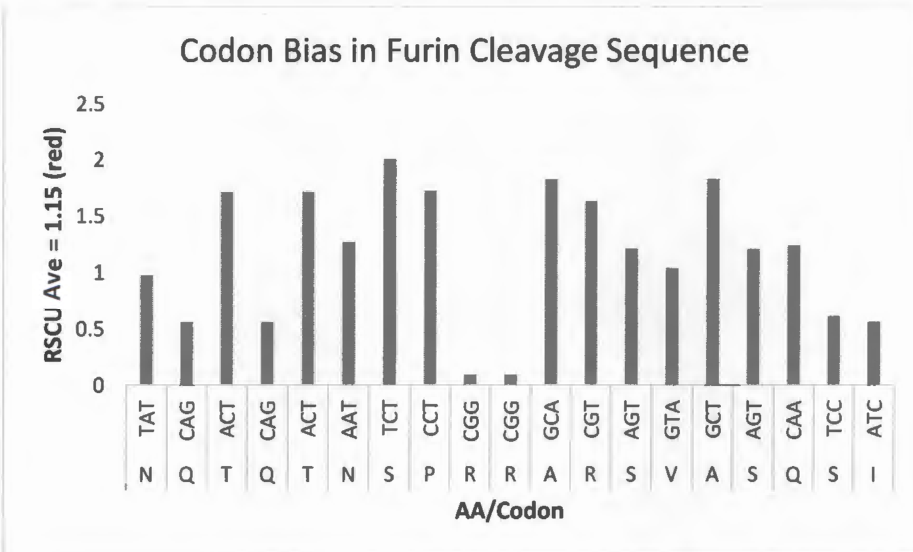
In this way, viruses in nature and scientists in the laboratory, with different goals and motivations, make distinguishing codon usage decisions which can sometimes provide a fingerprint of their source.

The Text-Table below contains the arginine codon usage for two populations, pooled data for SARS-CoV 2003 and related viruses and 13 Sars-CoV-2 human specimens from widely dispersed locations.

Codon	SARS-CoV 2003 and ten other evolutionary related viruses in the Nidovirales	SARS-CoV-2 from 13 Geo-locations
CGG	0.09	0.09
CGA	0.44	0.37
CGC	0.72	0.37
AGG	0.9	1.07
CGU	1.77	1.63
AGA	2.08	2.48

Since these values are of a type of multiplicative scale, they were fit using a log-normal distribution, which appears appropriate (although the sample size is small). Using the log mean and standard deviation and this distribution, the probability of finding a CGG codon is about 0.024. Assuming they are independent the probability of finding a CCG-CCG codon pair is effectively 0.024^2 or 0.00058. This is a likelihood of about one in 1700.

The following Figure shows the RSCU for the amino acids that comprise the new furin cleavage site in SARS-CoV-2. As one can see, the RSCU values are similar to each other with the exception of the RR dimer insert, which have a very low RSCU of 0.09.



The RSCU value for the CGG codon for R of 0.09 was taken from a 2004 paper of the RSCU for SARS-CoV 2003 and ten other evolutionary related viruses in the *Nidovirales* and is confirmed by 13 SARS-CoV-2 specimens obtained from diverse geographic locations. If one assumes that the RSCU observations are independent and that the probability distribution of these measurements is Gaussian (normal; a reasonable assumption), then one can calculate the probability of obtaining a result as small as 0.09. Removing the two 0.09 values, then the mean and standard deviation of the remaining values are 1.275 and 0.4992, respectively. Then the probability of a single 0.09 value is 0.0088. However, there are two 0.09 values. If we assume that these are independent findings, then the probability of both values being seen is 0.0088^2 or 7.7×10^{-5}. Using the RSCU of 0.2 from the Table above does not change the immense improbability of the usage of a CGGCGG codon pair in the wild.

Single Arginine CGG codon usage analysis suggests this will not be found in the wild.

The codon usage for SARS-CoV-2, like most coronaviruses studied, has a bias toward AT and away from GC nucleotides. The frequency of third position G use in CoV-2, for example, is 13%, 21%, 17%, and 16% for the spike protein, envelope, membrane, and nucleocapsid protein, respectively.

In that context, the scarcity of the CGG genome in SARS-CoV-2 and related coronaviruses, the relative synonymous codon usage, determined by the method of Behura and Severson,[87] was calculated and tabulated below. The color coding is blue for underutilized codons (RSCU < 1.0) and red for overutilized codons (RSCU > 1.0); light blue for RSCU values of 0.60 to 0.99 and

[87] https://www.ncbi.nlm.nih.gov/pubmed/22889422

light red for RSCU of 1.01 to 1.60. The highest RSCU usage of CGG is 1.21 in the membrane protein in the MERS virus but zero in SARS-CoV-2.

RSCU	SARS-CoV-2	Beta CoV Pangolin	SARS CoV	Bat SARS CoV	MERS CoV
Spike	0.29	0	0.19	0.08	0.25
Envelope	0	0	0	0	0
Membrane	0	0.35	0.74	0.24	1.21
Nucleocapsid	0.41	0.16	0.03	0.04	0.8

Looking at these five coronaviruses:

The largest structural protein of the coronaviruses is the spike protein, with 1273 amino acids. In SARS-CoV-2 there are 42 R residues, with only one RR dimer, the one in the insert that created SARS-CoV-2.

As a reminder none of these related coronaviruses have the 12-nucleotide insertion that forms the putative furin site in CoV-2. Interestingly, the pangolin coronavirus has no CGG residues in the spike protein. The significance of this is it makes the acquisition of this insert from pangolin by recombination impossible.

The smallest structural protein, the envelope protein, has 75 amino acids, including three R residues, but has no CGG codons in any of the related coronaviruses examined.

The SARS-CoV-2 membrane protein has 441 amino acids, 14 R residues and no CGG codons. Among related coronaviruses, this is the most unique finding of the four proteins for SARS-CoV-2 since the other four coronaviruses all utilize CGG to some extent in this protein. In the case of the MERS virus, this protein is the only occurrence in which this codon is overutilized.

The nucleocapsid protein has 418 amino acids and is responsible for packing the RNA genome. As expected for the role of R in protein-RNA interactions, it has 29 R residues and four RR dimers. None of the dimers use the CGGCGG sequence.

The nt usage of the 12-nt insert which forms the FCS cleavage site has a probability this sequence was selected for in the wild of one in 129,870.

A blast search was performed for the 12-nt inserted sequence and adjacent extensions and only the SARS-CoV-2 sequences were identified.

Shortening the search to just the two CGG-CGG codons was only slightly more fruitful. The Text-Table below shows the frequency of the middle half of the insert, CGGCGG, across the genomes of all seven known human coronaviruses, as well as a specimen bovine coronavirus and the bat and pangolin coronaviruses with greatest homology to SARS-CoV-2. Only a single example, outside of the Spike Protein gene, has been found.

Furin PBCS sequence	Beta Coronavirus	Total Arginine Dimers Anywhere	CGGCGG in Spike Protein *	CGGCGG Anywhere in genome *	CCGCCG Anywhere in genome
SRRKRRS	Human CoV-HKU1 GenBank: KF686346.1	12	0	0	0
KRRSRRA	Bovine CoV-Quebec GenBank: AF220295.1	12	0	0	0
PRRARSV	**SARS-CoV-2 Wuhan reference sequence GenBank: NC_045512.2**	**16**	**1; nt 23,606**	0	0
PRSVRS	MERS-CoV NCBI Reference Sequence: NC_019843.3	21	0	0	0
NRRSRGA	Human CoV-OC43 London/2011 GenBank: KU131570.1	16	0	0	0
None	Human CoV-229E GeneBank: KF514433.1	15	0	0	0
None	Human CoV NL63 NCBI Reference Sequence: NC_005831.2	9	0	0	0
None	SARS-CoV 2003 ZJ0301 from China GenBank: DQ182595.1	17	0	0	0
None	Bat coronavirus RaTG13 GeneBank: MN996532.1	11	0	1; nt 9394	0
None	Pangolin PCoV_GX-P4L GenBank: MT040333.1	10	0	0	0
	Total	139	1	0	0

* - Includes both in phase codons as well as out of phase, frameshift codons.

To understand what this means for the search for the zoonotic source for SARS-CoV-2, a statistical approach was taken. Using the data from the nine viruses other than SARS-COV-2 there was a single incidence of the CGGCGG found in the bat coronavirus. Assuming 10,000 codons per genome, the frequency of CGGCGG in coronaviruses can be estimated at 2 per 45,000 codons or 4 x 10^{-5}. Therefore, the frequency of finding the center half of the SARS-CoV-2 insert is very small. This is consistent with the strong bias in all coronaviruses to place an A/U nt in the third codon position.

The last column above, the presence of -CCG-CCG- in these coronaviruses was included because it is the hybridization sequence partner for the negative strand sequence, which arises during genome replication. This eliminates the possibility of a strand jumping event to generate a CGGCGG codon dimer.

A similar analysis for the spike protein gene can be done. Since there are no instances of CGGCGG in the spike protein genome, and the gene is 3819 nucleotides long, there are 636 pairs of codons Thus, over the 9 other viruses, there are 5724 pairs of codons and no cases of the CGGCGG pair. To calculate the upper bound on the probability of such a pair from these data, one can use the Poisson "Rule of Three", which yields a value of 3/5724 or 0.00052 with 95% confidence. Now examining the SARS-COV-2 genome, there was one instance of the pair in question out of 636 pairs. The probability of this happening if the true rate of this occurrence for a beta coronavirus is 0.00052 is 0.044. Obviously for smaller assumed rates of this occurrence, this would result in probabilities less than 0.044.

Since the 12-nt insert has been found nowhere in the coronavirus genomic universe, examining over 300,000 sequences and using the Poisson "Rule of Three" again, the upper bound on the frequency that it exists in nature is less than one in 100,000 with 95% confidence.

This observation in conjunction with the lack of finding the 12-nt sequence in any candidate zoonotic species makes unlikely a natural source for the virus. One line of investigation to establish a wild source for this infection would be to find a coronavirus strain with the 12-nt sequence somewhere in nature. The fact that 10 of the 12 nts are either G or C coupled, the documented bias against GC suggests this search would be futile.

Based on these analyses that demonstrate that the finding of a -CGG-CGG- codon pair in the furin site of CoV-2 is a highly improbable event, and using the conservative value of a one in 20 chance (the value for a p-value of 0.05), one can recalculate the likelihood of the choice between a zoonotic origin and a laboratory origin.

Subjective Discount Factor. 95% confidence (only a one in 20 chance this is wrong). Below is the calculation of the Bayesian adjustment.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.047	0.953
Negative predictive value of the absence of the -CGG-CGG- pair in any coronavirus in nature	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.047/9 = 0.005	
Normalize this step of analysis	0.005/(0.005 + 0.953) = 0.005	0.953/(0.953 + 0.005) = 0.995

Adjusted likelihood. Zoonotic origin (0.5%), laboratory origin (99.5%).

Evidence. Laboratory codon optimization uses CGG for laboratory insertions of arginine residues 50% of the time.

Codon optimization by recombinant methods (that is, to bring a gene's synonymous codon use into correspondence with the host cell's codon bias) has been widely used to improve cross-species expression of protein.

Though the opposite objective of reducing expression by intentional introduction of suboptimal synonymous codons has not been extensively investigated, isolated reports indicate that replacement of natural codons by rare codons can reduce the level of gene expression in different organisms. For example, one approach to vaccine development is to create an attenuated virus which comprises a modified viral genome containing nucleotide substitutions engineered in multiple locations in the genome, wherein the substitutions introduce synonymous de-optimized codons.

In US Patent 9,476,032[88] titled, "Attenuated viruses useful for vaccines," they state: "In one high-priority redesigned virus, most or all Arg codons are changed to CGC or **CGG** (the top two frequent human codons). This does not negatively affect translation." The patent contains numerous codon usages optimized for vaccine production, including the SARS-CoV virus, and in fact they use the CGG-CGG codon pair 45 times.

Beginning with a paper in 2004,[89] one motivation for codon-optimized SARS genomes is stated here: "The gene encoding the S protein of SARS-CoV contains many codons used infrequently in mammalian genes for efficiently expressed proteins. We therefore generated a codon-optimized form of the S-protein gene and compared its expression with the S-protein gene of the native viral sequence. S protein was readily detected in HEK293T cells transfected with a plasmid encoding the codon-optimized S protein."

Since that time, human optimized codons have been frequently used for coronavirus research, mostly in gain-of-function experiments. In that context the "molecular fingerprint" of CGG for R is one of those common laboratory reagent gene manipulators.

Other examples:

Examples of the use of CGG codon **for arginine in coronavirus research**	**Reference**
SARS was genetically modified to improve ACE2 binding using "human optimized" codons, like CGG for arginine, to grow better in the laboratory. The strains were more infective.Preparation of SARS-CoV S protein pseudotyped virus. "The full-length cDNA of	Wu, K. et al. Mechanisms of Host Receptor Adaptation by Severe Acute Respiratory Syndrome

[88] http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=9476032.PN.&OS=PN/9476032&RS=PN/9476032
[89] https://www.ncbi.nlm.nih.gov/pubmed/15367630

the SARS-CoV S gene was optimized according to human codon usage and cloned into the pCDNA3.1(+) vector (Invitrogen). The resulting "humanized" S sequence was identical with that of strain BJ01 at the amino acid level."	Coronavirus. J Biol Chem. 2012 Mar 16; 287(12): 8904–8911.
Predictions of future evolution of a virus are a difficult, if not completely impossible, task. However, our detailed structural analysis of the host receptor adaptation mutations in SARS-CoV RBD has allowed us to predict, design, and test optimized SARS-CoV RBDs that may resemble future evolved forms of the virus. "RBD might evolve into the human-optimized form by acquiring two mutations at the 442 and 472 position." SARS-CoV-2 acquired the mutation at position 472.	Fang Li. Receptor recognition and cross-species infections of SARS coronavirus. Antiviral Res. 2013 Oct; 100(1): 246–254.
Plasmid encoding a codon-optimized form of the SARS-CoV S protein of the TOR2 i	Wenhui Li, Chengsheng Z, et al., Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 2005 Apr 20; 24(8): 1634–1643.
The gene encoding the S protein of SARS-CoV contains many codons used infrequently in mammalian genes for efficiently expressed proteins. We therefore generated a codon-optimized form of the S-protein gene and compared its expression with the S-protein gene of the native viral sequence. S protein was readily detected in HEK293T cells transfected with a plasmid encoding the codon-optimized S protein (Fig. (Fig.1).1). No S protein was detected in cells transfected with a plasmid encoding the native S-protein gene.	Moore, MJ, Dorfman, T. Retroviruses Pseudotyped with the Severe Acute Respiratory Syndrome Coronavirus Spike Protein Efficiently Infect Cells Expressing Angiotensin-Converting Enzyme 2. J Virol. 2004 Oct; 78(19): 10628–10635.
Published in 2019 by **Dr. Zhengl-Li Shi,** entitled "Origin and evolution of pathogenic coronaviruses," reviews genetic optimized SARS viruses using human codons.	Cui, J, Fang, L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019; 17(3): 181–192.
In 2006, Montana scientists put a synthetic furin cleavage site into a SARS coronavirus by adding an R residue at position R667. They write: "We show that furin cleavage at the modified R667 position generates discrete S1 and S2 subunits and potentiates membrane fusion activity." Mutations were introduced by using	Follis, KE, York, J, Nunberg, JH. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell–cell fusion but does not affect virion entry. Virology 350 (2006) 358–369

QuikChange mutagenesis (Stratagene)[90]	
Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein is the title of a paper that shows that the expression of spike protein in vitro was greatly increased by expression cassette optimization.	Zhia, Y, Kobinger, GP, Jordan, H, et al. Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein
As for the human clec4C_1 and mouse clec14A, they showed very similar profiles with spike genes, especially with bat SARS-CoV, in the arginine coding groups, showing the high RSCU values over 2.50 in AGA.	Ahn,I, Jeong, B-J, Son, HS. Comparative study of synonymous codon usage variations between the nucleocapsid and spike genes of coronavirus, and C-type lectin domain genes of human and mouse. Experimental & Molecular Medicine volume 41, pages746–756, 2009.

One relevant paper,[91] in which arginine residues were being inserted into bovine herpesvirus-1, used primers to create RR dimers with nine separate -CGG-CGG- codon pairs. as testament to their broad use in the Wuhan Institute of Virology laboratory.

Scientists from the Wuhan Institute of Virology provided the scientific community with a technical bulletin on how to make genetic inserts in coronaviruses and proposed using the very tool that would insert this CGGCGG codon.

A Technical Appendix[92] entitled, "Detailed methods and primer sequences used in a study of genetically diverse filoviruses in Rousettus and Eonycteris spp. bats, China, 2009 and 2015, by Yang, Xinglou & Zhang, Yunzhi & Jiang, Ren-Di & Guo, Hua & Zhang, Wei & Li, Bei & Wang, Ning & Wang, Li & Rumberia, Cecilia & Zhou, Ji-Hua & Li, Shi-Yue & **Daszak, Peter** & Wang, Lin-Fa & **Shi, Zheng-Li.** (2017), from the Wuhan Institute of Virology identifies primer sequences for doing genetic experiments in coronaviruses and identifies CGG containing primers when a R amino acid is being inserted.

[90] Since the codon usage here was not reported I contacted Professor Nunberg to inquire which arginine codons were used. He replied: "Unfortunately, those files have all been archived and access to the nt sequences would involve considerable digging. If it is useful to you, I typically choose codons that are more frequent in highly expressed human proteins."
[91] From the Wuhan Institute of Virology; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7125963/
[92] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382765/

Given that there are two codons of six possibilities that are used in codon optimization, CGG and CGC, the finding of a CGG pair would have a likelihood of happening by chance of (2/6) times (2/6) or one in nine.

Subjective Discount Factor: 80% (this has a probability of being wrong one in five times). This is arbitrary. The calculation to make this adjustment in likelihood is shown here:

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.005	0.995
This is the outcome expected 8 of 9 times if this is codon optimization		0.88
Reduced by 80% confidence		0.88 x 0.8 = 0.704
Impact of this evidence		Increases the likelihood of LO by 70.4 divided by 29.6 or 2.378.
Impact of evidence calculation		0.995 x 2.378 = 2.37
Normalize this step of analysis	0.005/(2.37 + 0.005) = 0.002	2.37/(0.005 + 2.37) = 0.998

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: SARS-CoV-2 Spike Protein is Highly Optimized for ACE2 Binding and Human Cell Infectivity, a Finding that is Inconsistent with Natural Selection but is Consistent with Laboratory Creation

Summary:

- Andersen et al.[93] hypothesized that if the CoV-2 interaction with the human ACE2 was apparently "not ideal," it was evidence that CoV-2 arose by natural selection.

- The alternative hypothesis would be that a finding that CoV-2 was optimized for ACE2 binding and human infection from the initial infection would be evidence of laboratory creation.

- Andersen relied on a paper for the "not ideal" interaction that relied on a computer algorithm rather than laboratory data, was qualitative in nature, sampled only five amino acids or 0.45% of the interaction region, and was over-interpreted.

- The analysis of the Baric et al. paper cited by Andersen as evidence the interaction was not ideal was reexamined, and it was concluded that Andersen had over-interpreted the paper. The paper was a computer simulation study of only 5 of 201 amino acids in the CoV-2-ACE2 interaction region. Only one of the five amino acids discussed was said to be inferior to the equivalent amino acid in SARS-CoV-1; the remainder were either positive or neutral with respect to binding.

- More recently, Baric has clarified his thoughts concerning the CoV-2 ACE2 receptor binding interaction. In a December 31, 2020 *New England Journal of Medicine* paper[57] he wrote: "Early zoonotic variants in the novel coronavirus SARS-CoV that emerged in 2003 affected the receptor-binding domain (RBD) of the spike protein and thereby enhanced virus docking and entry through the human angiotensin-converting–enzyme 2 (hACE2) receptor. **In contrast, the spike-protein RBD of early SARS-CoV-2 strains was shown to interact efficiently with hACE2 receptors early on.**" [emphasis added.]

- A comprehensive, laboratory-based, and quantitative paper by Starr et al. of all 201 amino acids in the receptor binding region, not just five amino acids, was examined. Fully 99.6% of all of the possible 3819[94] amino acid substitutions were tested for their effect on CoV-2 binding to ACE2. Only 21 substitutions of the 3819 improved ACE2 binding. Therefore, CoV-2 has been optimized for human ACE2 binding in 99.45% of the possible amino acids in its Spike Protein interaction region.

---

[93] https://www.nature.com/articles/s41591-020-0820-9

[94] There are 201 amino acids in the residue 331 to 531 interaction region and so 201 times the 19 possible alternative amino acids not found in CoV-2 equals 3819.

- To support this finding, Starr also made an examination of 31,570 CoV-2 sequences from human infections, looking for the 21 substitutions that had been shown to improve CoV-2 binding in the above in vitro laboratory experiments. Among the 31, 570 CoV-2 cases, they failed to find even a single case in which there was an amino acid substitution that improved binding at the time of writing this analysis.[95]

- Based on Andersen's hypothesis and its alternative, SARS-CoV-2 is fully optimized for interaction with the human ACE2 receptor and was at the time of the first patient. There is no evidence of an evolving SP binding region, as was seen with SARS-CoV-1. This is consistent with a laboratory optimized coronavirus which entered the human population fully evolved.

<u>Analysis</u>

Quote from Andersen: "While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal (reference 7) and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding (references 7,11).

Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 that permits another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation."

The apparent **<u>hypothesis</u>** for the above conclusion is:

"If the SARS-CoV-2 (CoV-2) Spike Protein interaction with the ACE2 receptor is not maximized, then it is evidence that the interaction is the product of natural selection and not purposeful (laboratory) manipulation."

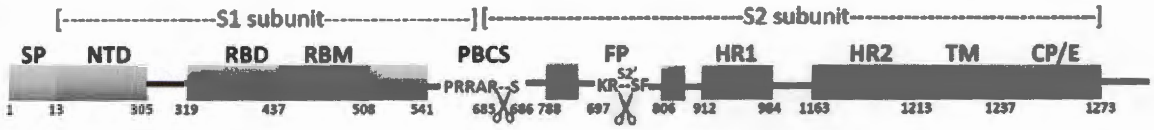
This would lead to an **<u>alternative hypothesis</u>**:

"If the CoV-2 Spike Protein interaction with the ACE2 receptor is maximized, then it is evidence that the interaction *was* the product of purposeful (laboratory) manipulation."

Background.

The Spike Protein (SP) structure and its functional domains are shown in this Figure. The S1 subunit is the initial host interaction portion while the S2 is the post-binding portion responsible for initiating host cell entry, with HR1, HR2, and TM being responsible for breaching the host cell membrane. Allowing viral RNA to enter the cell.

[95] The recent finding of the N501Y variant, first in the UK, and now spreading globally, is evidence of the power of this analysis. N501Y is one of only five potential substitutions in the Starr analysis that had a major effect in improving ACE2 binding.



The interaction of the SP portions which interact with the ACE2 of the host cell, which begins the internalization, infectious process, are contained in the Receptor Binding Domain (RBD) and to a lesser extent the Receptor Binding Motif (RBM), specifically residues 331 to 531. Herein, residues 331 to 531 are called the "interaction region."

__Evidence given by Andersen:__

Reference 7 in the Andersen paper above is a Ralph Baric paper[96] from early in the pandemic (submitted January 22, 2020) and examines five key residues in the receptor binding domain of the Spike Protein (SP) and whether they are "ideal" for interacting with the ACE2 of human cells. The entire paper is based on computer calculations or prior laboratory work but importantly does not do any new "wet" lab work with CoV-2.

Baric et al. had previously identified five amino acid residues that are important for SP-ACE2 interaction. Using the amino acid numbers of CoV-2, these amino acids are: 455, 486, 493, 494, and 501. Baric opines that the most critical residues are 493 and 501 and the next most important residues are 455, 486, and 494. The authors then discuss each amino acid in turn:

__Residue 493__: "Gln493 in 2019-nCoV RBD is compatible with hot spot 31, suggesting that 2019-nCoV is capable of recognizing human ACE2 and infecting human cells." In this analysis, 4 of the 20 amino acids are probed.

__Residue 501:__ "This analysis suggests that 2019-nCoV recognizes human ACE2 less efficiently than human SARS-CoV (year 2002) but more efficiently than human SARS-CoV (year 2003). Hence, at least when considering the ACE2-RBD interactions, 2019-nCoV has gained some capability to transmit from human to human."

Direct binding evidence has shown that this statement is misleading, and CoV-2 binds the ACE2 receptor about ten-times better than SARS-CoV (year 2002).[97] In this analysis 3 of the 20 amino acids are probed.

__Residues 455, 486, and 494:__ First, Baric et al. state: "Leu455 of 2019-nCoV RBD provides favorable interactions with hot spot 31, hence enhancing viral binding to human ACE2."

Next, they state: "Phe486 of 2019-nCoV RBD provides even more support for hot spot 31, hence also enhancing viral binding to human ACE2." Importantly, they also talk about their own laboratory work on an "optimized" receptor binding domain and state: "Leu472 of human and

---

[96] https://jvi.asm.org/content/94/7/e00127-20
[97] https://www.cell.com/action/showPdf?pii=S0092-8674%2820%2931003-5 ; https://www.nature.com/articles/s41586-020-2179-y ; https://www.sciencedirect.com/science/article/pii/S0092867420302622 ; https://science.sciencemag.org/content/367/6483/1260

civet SARS-CoV RBDs provides favorable support for hot spot 31 on human ACE2 through hydrophobic interactions with ACE2 residue Met82 and several other hydrophobic residues **(this residue has been mutated to Phe472 in the optimized RBD)**." [emphasis added.]

Finally, they state: Ser494 in 2019-nCoV RBD still provides positive support for hot spot 353, but the support is not as favorable as that provided by Asp480. Overall, Leu455, Phe486, and Ser494 of 2019-nCoV RBD support the idea that 2019-nCoV recognizes human ACE2 and infects human cells."

In this analysis they probe 3 of 20 amino acid residues for position 480, 4 of 20 for position 486, and 4 of 20 for position 442.

As shown in the Figure below from the Baric paper, the in vitro designed, optimized human SP (red arrow) had the amino acid residues F, F, N, D, and T at these five key residues. Since CoV-2 was identical in only one of these five it was not "optimal" and, according to Andersen, it therefore was not laboratory derived.

B

Virus	Year	442	472	479	480	487
SARS - human	2002	Y	L	N	D	T
SARS - civet	2002	Y	L	K	D	S
SARS - human/civet	2003	Y	P	N	G	S
SARS - civet	2005	Y	P	R	G	S
SARS - human	2008	F	F	N	D	S
Viral adaption to human ACE2		F > Y	F > L > P	N = R >>> K	D > G	T >>> [illegible]
Optimized - human	In vitro design	F	F	N	D	T
Viral adaptation to civet ACE2		Y > F	P = L > F	R > K = N	G > D	T > S
Optimized - civet	In vitro design	Y	P	R	G	T
SARS - bat	2013	S	F	N	D	N
2019-nCoV – human	2019	L (455)	F (486)	Q (493)	S (494)	N (5[illegible])

Conclusion from the above paper: by examining five amino acid residues of the 200 residues encompassing the interaction region, and calculating the expected interaction of a total of 18 of the 4000 possible residues or 0.45% of all possibilities, they conclude CoV-2 can infect human cells, but is not optimized to do so. This data was twisted by Andersen to show 'strong evidence' of natural selection.

An alternative and comprehensive analysis in another paper:[98]

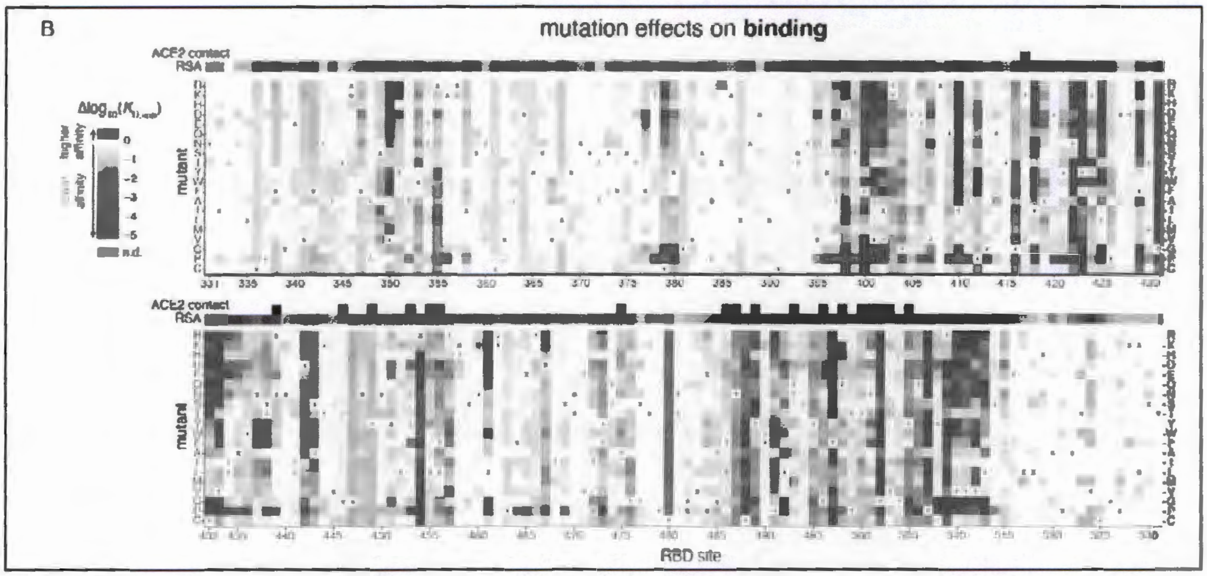
The receptor binding domain (RBD) of the CoV-2 SP is included in residues 331 to 531, a 201 amino acid sequence, of the SP. To examine the effect of each and every amino acid in each and every position, all 19 different amino acids were changed into all 201 positions of the RBD to the extent possible. Out of a total potential of 3819 different single amino acid variants, the scientists

[98] https://www.cell.com/action/showPdf?pii=S0092-8674%2820%2931003-5

were able to create 3804 of the potential variants or 99.6% of the possible variants. It is probable that the variants with the 0.4% amino acid substitutions could not be made for one reason or another. These 3804 were then tested for binding to the human ACE2. Finally, the RBD from SARS-CoV-1 also was tested.

The Figure below is the result of the experiment. Starting with amino acid 331 and ending with amino acid 531, the amino acids that were changed are in vertical columns and are color coded. Shades of brown are amino acid substitutions that reduce ACE2 binding affinity and blue are amino acid substitutions that improve binding, in all cases compared to the 'native' CoV-2 SP sequence. White is the color of a neutral substitution which neither enhances nor diminishes binding. Only the dark blue substitutions provide a strong improvement in ACE2 binding. There is a black square along the top row that denotes amino acids in the SP that interact with the ACE2 protein. Unlike in the Baric analysis above, in which only five amino acids were considered, this group of 19 amino acids provide a more complete interaction picture.

The first overarching observation is that most amino acid substitutions among the 201 amino acids are negative; while a large number are neutral. The fact that the vast majority of amino acid substitutions do not provide an improved ACE2 interaction is clear evidence that the CoV-2 SP interaction region is not newly evolved to the human ACE2 but arrived in the first patient having been "trained" to invade and kill human cells.



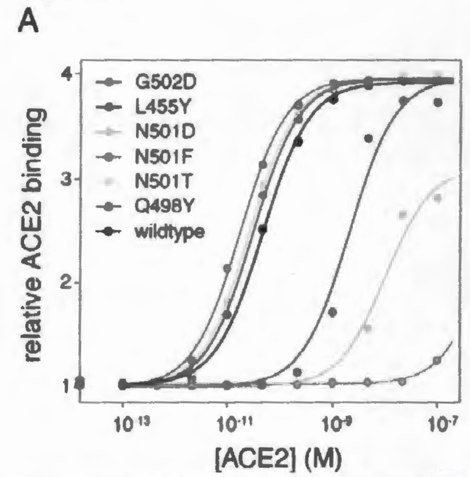
There are three levels of improved binding as designated by dark blue, medium blue, and pale blue. Out of the 3804 variants tested, there are 4 dark blue substitutions or 0.11% and 17 medium blue or 0.45%. According to the paper, the binding effect of the light blue could not be measured as different from the native sequence.

The conclusion of this comprehensive work is the demonstration that for 99.45% of the amino acids in the 201 amino acid interaction region, the CoV-2 choice is optimized, where any substitution is either detrimental or, at best, neutral with respect to the first step of CoV-2 entry to human cells, the binding step to the ACE2 receptor.

How much could CoV-2 binding be improved or made worse by substitutions during the human-to-human transmission of the pandemic?

The Figure 4 below, taken from the paper, shows that the three best amino acid substitutions have only a slight effect on the binding curve (Black is wildtype; curves to the left are better binding; curves to the right are worse binding). This is further evidence that CoV-2 is an optimized form of the original virus.

Figure 4



The authors also concluded that Anderson et al. was wrong: "An initially surprising feature of SARS-CoV-2 was that its RBD tightly binds ACE2 despite differing in sequence from SARS-CoV-1 at many residues that had been defined as important for ACE2 binding by that virus (Andersen et al., 2020; Wan et al., 2020)."

In fact, multiple studies have shown that CoV-2 binds ACE2 better than SARS-CoV-1, contradicting Andersen.

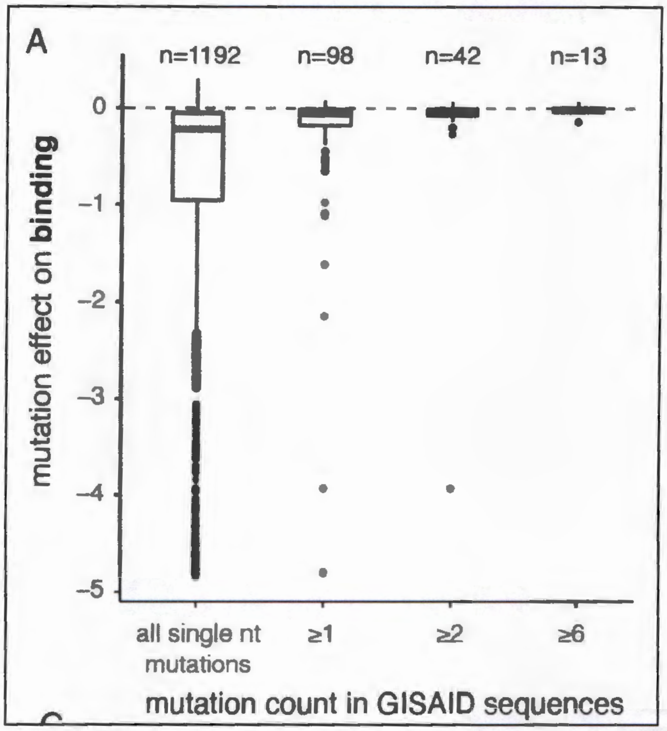
Is there evidence that CoV-2 in human circulation has mutations that enhance ACE2 binding?

Another measure of whether CoV-2 is optimized for human infection is to see if Spike Protein mutations have arisen during the pandemic that improve binding of the virus to the ACE2 receptor or if the SP amino acids are ideal from the very first human patient.

The Starr paper addressed this issue as well. A total of 31,570 human sequences were analyzed to see if any of the 21 amino acid substitutions from the binding experiments (or any other for that matter) were being selected for. That is, if there is any evidence of evolutionary pressure to improve SARS-CoV-2 infectivity.

Below is Figure 8 of the Starr paper. Of the 31,570 sequences, all mutations in the receptor interaction region were analyzed for their effect on ACE2 binding. The data below are for all examples of a single nt mutation (1192), two mutations (98), 3-5 mutations (42), and six or more (13) and the effect the mutation would have on ACE2 binding. The logarithmic scale has the

wildtype CoV-2 as 0 and each negative integer is a 10-fold reduction in affinity. Shockingly, there is not a single mutation that is above the 0 line, which would be an improved affinity for the ACE2 receptor. All of the mutations lower the receptor affinity.



Here are the results, in the words of Starr:

"Our discovery of multiple strong affinity-enhancing mutations to the SARS-CoV-2 RBD raises the question of whether positive selection will favor such mutations, since the relationship between receptor affinity and fitness can be complex for viruses that are well-adapted to their hosts (Callaway et al., 2018; Hensley et al., 2009; Lang et al., 2020). Strong affinity-enhancing mutations are accessible via single-nucleotide mutation from SARS-CoV-2 (Figure S8C), but **none are observed among circulating viral sequences in GISAID** (Figure 8A), and **there is no significant trend for actual observed mutations to enhance ACE2 affinity more than randomly drawn samples of all single nucleotide mutations** (see permutation tests in Figure S8D). **Taken together, we see no clear evidence of selection for stronger ACE2 binding, consistent with SARS-CoV-2 already possessing adequate ACE2 affinity at the beginning of the pandemic."** [emphasis added.]

It is striking that the authors, in observing the complete absence of any evidence for stronger ACE2 binding in over thirty thousand cases, would describe this as evidence of "adequate ACE2 affinity" and not as an exceptional finding of "optimized ACE2 affinity." Of course, calling the SP affinity exceptional from the beginning of the pandemic would beg the question of a laboratory derived virus.

Returning to the initial hypotheses, since the 3804 possible amino acids at the receptor interaction region of CoV-2 are 99.45% optimized for ACE2 binding, and there is not a single